LIETUVIŲ KALBOS INSTITUTAS

DAIVA ŠVEIKAUSKIENĖ

LIETUVIŲ KALBOS GRAMATIKOS KOMPIUTERIZAVIMAS

MOKSLO STUDIJA

Vilnius
2022

Patvirtinta spausdinti Lietuvių kalbos instituto Mokslo tarybos 2022 m. gruodžio 1 d. posėdyje, protokolas Nr. MT-6.

RECENZENTAI:

Prof. dr. Albinas DRUKTEINIS (Klaipėdos universitetas)

Doc. dr. Pijus KASPARAITIS (Vilniaus universitetas)

Doc. dr. Algirdas LAUKAITIS (Vilniaus Gedimino technikos universitetas)

Pataisytas leidimas

Bibliografinė informacija pateikiama
Lietuvos integralios bibliotekų informacinės
sistemos (LIBIS) portale *ibiblioteka.lt*.

DOI doi.org/10.35321/e-pub.44.lietuviu-gramatikos-kompiuterizavimas

ISBN 978-609-411-327-7

TURINYS

PRATARMĖ

Apdorojant kalbą kompiuteriu, iškyla daug problemų ir klausimų, kurie anksčiau kalbininkų nebuvo aptarti, nes jie paprasčiausiai nebuvo aktualūs. Žmogus, gavęs dalį informacijos apie pačius bendriausius kalbos dalykus, pateiktus spausdintose gramatikose, gali pats susigeneruoti trūkstamus ar akivaizdžiai neišreikštus duomenis apie žodžius ir sakinius, remdamasis gimtosios kalbos jausmu. Popieriuje spausdintų gramatikų užteko, kol jų vartotojas buvo žmogus. Atsiradus naujam vartotojui – kompiuteriui – iškilo daug problemų, susijusių su kalbinės informacijos perteikimu jam. Kompiuteris negali pats pasiskaityti išleistų gramatikos knygų ir iš jų pasiimti informacijos apie kalbą. Tai turi būti pateikta formaliai, t. y. jam suprantamu būdu. Buvo sukurta daug įvairių metodų, kaip kompiuteriui perduoti žinias apie kalbą, ir dalis jų bus aptarta šioje mokslo studijoje tiek apžvelgiant kitų mokslininkų atliktus darbus, tiek pateikiant ir savus mokslinius tyrimus.

Šiame darbe aprašomi autorės jau anksčiau publikuoti ir toliau tęsiami lietuvių kalbos gramatikos tyrimai, kurių poreikis iškyla kompiuterizuojant kalbą. Kuriant *Lietuvių kalbos gramatikos informacinę sistemą*, siekiama išvengti trūkumų, pastebėtų jau atliktuose darbuose. Vienas jų – nepakankama žodžių apimtis tiek tekstynuose, tiek žodynuose. Kad būtų galima sugeneruoti visus teoriškai įmanomus lietuvių kalbos žodžių vedinius ir dūrinius, reikia išanalizuoti darybines morfemas. Todėl Lietuvių kalbos institute (toliau – LKI) buvo atlikta išsami priešdėlių analizė (Šveikauskienė 2015a). Kita problema iškilo dėl nevienodo kalbos dalių traktavimo skirtinguose šaltiniuose. Kadangi *Lietuvių kalbos gramatikos informacinėje sistemoje* pasirinktas dalyvio statusas nesutampa su pateikiamu akademinėje *Lietuvių kalbos gramatikoje*, šioje mokslo studijoje, siekiant pagrįsti tokį pasirinkimą, visas poskyris parašytas remiantis straipsniu apie dalyvį (Šveikauskienė 2018). Dar vienas straipsnio pagrindu parengtas poskyris – apie lietuvių kalbos skaitmeninę gramatiką (Šveikauskienė 2019a). Likusioje studijos dalyje panaudota medžiaga ir iš kitų straipsnių (Šveikauskienė 2013, 2014, 2015b, 2016, 2017, 2019b, 2021).

Dėkoju Lietuvių kalbos instituto Bendrinės kalbos tyrimų centro vadovei dr. Jurgitai Jaroslavienei ir šio skyriaus darbuotojams dr. Daivai Murmulaitytei, dr. Ritai Miliūnaitei, dr. Loretai Semėnienei, dr. Mindaugui Strockiui, dr. Veslavai Čižik-Prokaševai, dr. Aurelijai Tamulionienei, dr. Jolitai Urbanavičienei, dr. Aurelijai Gritėnienei, dr. Anželikai Gaidienei bei kitiems Lietuvių kalbos instituto darbuotojams dr. Petrui Skirmantui, Vytautui Zinkevičiui, perskaičiusiems studijos rankraštį ir prisidėjusiems prie jos vertingomis pastabomis ir patarimais.

Už pagalbą, rengiant visus mano tekstus vokiečių kalba ir šios studijos santrauką, noriu padėkoti akad. prof. habil. dr. Grasildai Blažienei. Taip pat už pagalbą dėl vokiečių kalbos dėkoju dr. Dariui Ivoškai.

Už vertingas pastabas dėl lietuvių kalbos gramatikos dėkoju recenzentui prof. dr. Albinui Drukteiniui; taip pat dėkinga recenzentui doc. dr. Pijui Kasparaičiui už nepaprastai atidų mano studijos peržiūrėjimą ir pateiktus pasiūlymus; dėkoju recenznetui doc. dr. Algirdui Laukaičiui už išsakytas mintis apie statistinių metodų perspektyvą.

1. ĮVADAS

Šios mokslo studijos tema yra iš mokslų sandūros (lingvistikos ir informatikos) srities, todėl ir klausimai bus aptariami dvejopai – iš kalbinės ir iš kompiuterinės pusės.

1.1. Bendrosios pastabos

Atsiradus kompiuteriams, jie pradėjo skverbtis beveik į visas žmogaus gyvenimo sferas, kalbos – ne išimtis. Pasaulyje jau daug padaryta jas kompiuterizuojant. Nemažai darbų atlikta ir Lietuvoje. Vis dėlto kol kas nėra išsamesnės apžvalgos apie Lietuvos mokslininkų nuveiktus lietuvių kalbos gramatikos kompiuterizavimo darbus.

Mokslo studijos tikslas – aprašyti kitų mokslininkų atliktus darbus, susijusius su lietuvių kalbos gramatikos kompiuterizavimu, kompiuterizuotos gramatikos panaudojimu ar gramatinių duomenų pateikimu kompiuterine forma, bei pristatyti šios mokslo studijos autorės tyrimus, atliktus kuriant skaitmeninę gramatiką ir gramatikos informacinę sistemą.

Mokslo studijos uždaviniai. Svarbiausios sritys, kuriose galima įžvelgti lietuvių kalbos gramatikos sąsają su kompiuteriais, yra šios: tekstynų anotavimas, analizatorių kūrimas abiem gramatikos dalims – tiek morfologijai, tiek sintaksei, pati skaitmeninė gramatika ir gramatikos informacinė sistema.

Pagrindiniai mokslo studijos uždaviniai:

1) aprašyti tekstynų anotavimo dalį, kurioje nurodomi gramatiniai duomenys – tiek morfologiniai, tiek sintaksiniai;
2) apžvelgti atliktus morfologinių analizatorių kūrimo darbus ir pasiekimus morfemikos srityje;
3) pateikti sintaksinių analizatorių veikimo metodikos aprašymą ir darbo rezultatų analizę;
4) parodyti lietuvių kalbos skaitmeninės gramatikos bandomąjį pavyzdį, išsamiai aprašant skaitmeninių gramatikų kūrimo metodiką;
5) aptarti pradėtą kurti *Lietuvių kalbos gramatikos informacinę sistemą*.

Mokslo studijoje kiekvienam uždaviniui numatyta po skyrių.

Naujumas ir aktualumas. Šioje mokslo studijoje pirmą kartą pateikiama apibendrinta visų sričių, susijusių su lietuvių kalbos gramatikos kompiuterizavimu, apžvalga ir aprašomi anksčiau kalbininkų neaptarti duomenys apie lietuvių kalbos gramatiką. Jie buvo pastebėti gramatikos informacinės sistemos kūrimo metu (6.3 skyrius: „LIGIS išgauta informacija netyrinėtais lietuvių kalbos gramatikos klausimais“).

Aktualu yra išsamiau aptarti šiuo metu jau sukurtų lietuvių kalbos apdorojimo įrankių darbo rezultatus ir atkreipti visuomenės dėmesį į tai, kad naujausios, tarptautiniu mastu naudojamos kalbos technologijos dažnai pateikia netikslią informaciją, t. y. analizatoriai neteisingai nurodo lietuvių kalbos žodžių morfologinius požymius ir sudaro sakinių sintaksines struktūras su klaidomis. Todėl būtina kurti savus aukšto tikslumo ir patikimumo lietuvių kalbos kompiuterinio apdorojimo įrankius.

Metodika. Siekiant atskleisti lietuvių kalbos gramatikos kompiuterizavimo ypatumus, iš pradžių apžvelgiami analogiški kitų šalių darbai. Po to pateikiami pagal tuos pačius metodus atlikto lietuvių kalbos kompiuterinio apdorojimo rezultatai. Pabaigoje aprašomi bandymai kompiuterizuoti gramatiką išvengiant labiausiai išryškėjusių trūkumų jau atliktuose darbuose (nepakankama žodžių apimtis tekstynuose ir žodynuose, pertekliniai žodžiai kai kuriuose analizatoriuose ir kt.) ir kurti sistemas, pateikiančias labai aukšto tikslumo ir patikimumo gramatinę informaciją, t. y. sistemas, kuriose netoleruojamos klaidos. Darbe taikyta atrankos metodika, t. y. tyrimui atlikti atrinkti tie probleminiai atvejai, kurie paaiškėjo kompiuterizuojant kalbą. Darbas yra aprašomojo, analitinio, lyginamojo pobūdžio.

Tikslinės grupės. Ši humanitarinių mokslų, filologijos (H 04) srities mokslo studija daugiausia skiriama kalbininkams, planuojantiems savo darbuose naudoti informacines technologijas, taip pat plačiajai visuomenei, besidominčiai naujausiomis kalbų apdorojimo metodikomis. Todėl daugiau vietos skiriama kalbiniams klausimams, ne taip detaliai analizuojami kompiuterio veikimo principai bei algoritmai, kurie naudojami kompiuterizuojant lietuvių kalbos gramatiką. Jie aptariami tik pačiais bendriausiais bruožais. Pagrindinis dėmesys sutelkiamas į kompiuteriu atliekamų darbų rezultatus, jų aprašymą ir analizę. Informaciją apie gramatikos kompiuterizavimo darbus tiek pasaulyje, tiek Lietuvoje siekiama pateikti populiariai ir visiems suprantamai.

Kiekviename mokslo studijos skyriuje pateikiama šiek tiek teorijos ir analogiški darbai, atlikti kitose šalyse. Plačiau aptariami kai kurie informatikos srities klausimai, kad būtų galima aiškiau suprasti kalbos kompiuterizavimą.

Dėl mokslo studijoje vartojamų terminų. Terminas *gramatika* šioje studijoje suprantamas taip, kaip ji apibrėžta akademinėje *Lietuvių kalbos gramatikoje*: „Gramatika susideda iš dviejų skyrių: morfologijos ir sintaksės“ (Ulvydas ir kt. 1965: 10).

Nors Seimo nutarime „Dėl lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 metų gairių patvirtinimo“ (1 interneto nuoroda[1]) vartojami terminai *mašininis vertimas* ir *natūralioji kalba*, šioje studijoje pasirinkti *automatinio vertimo* ir *tautų kalbų* terminai.

Toks sprendimas grindžiamas *Tarptautinių žodžių žodyne* pateiktu pavyzdžiu prie žodžio *automatinis*: „[...] vykstantis arba veikiantis visai ar iš dalies pats, žmogui tiesiogiai nedalyvaujant, pvz., įrenginys, ginklas, procesas (valdymas, vertimas)“ (Kvietkauskas 1985: 55). Terminas *mašininis* naujausiame *Tarptautinių žodžių žodyne* apibrėžiamas taip: „[...] susijęs su mašina, skirtas mašinai, pvz., mašininis žemės dirbimas, mašininė alyva, mašininė imitacija“ (Vaitkevičiūtė 2007: 690). Prie šio termino pateikta nuoroda į *mašina 1*: „Įrenginys, turintis vieną ar kelis mechanizmus energijai transformuoti arba mechaniniam darbui atlikti, pvz., energetikos mašina – mašina, paverčianti bet kurios rūšies energiją mechanine arba atvirkščiai, darbo mašina – mašina, keičianti medžiagos (darbo objekto) formą, savybes, būseną, padėtį“ (Vaitkevičiūtė 2007: 689). Kiti mokslininkai kalbos kompiuterinio apdorojimo procesus taip pat vadina *automatiniais*, pvz., aprašant tekstynų anotavimą naudojantis V. Zinkevičiaus sukurta programine įranga, sakoma: „Automatinis morfologinis tekstynų anotavimas“. Ir kiti terminai – *automatinė sintaksinė ir semantinė analizė*, *automatinis morfologinis vienareikšminimas* (Rimkutė, Daudaravičius 2007: 30) – parodo, kad *automatinis* tiksliau nusako kompiuterių atliekamą darbą. Be to, šį terminą vartoja Europos Komisijos vertimo raštu specialistai pačiose naujausiose savo publikacijose, pvz., 2021-12-01 vykusio trečiojo ELRC seminaro Lietuvoje metu

[1] Prieiga internete: https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/911407f20ee911ebbedbd456d2fb030d?positionInSearchResults=0&searchModelUUID=a5713c95-b943-4db5-81c0-84b7c1d74208&fbclid=IwAR0IS_88QUeoq_Cv_7XaLj0mIWoHqu7JOarz1AaCn4MfqIYogOQafELXOmk [žiūrėta 2022-11-22].

Vilmantas Liubinas skaitė pranešimą tema „CEF automatinio vertimo platforma“ (Liubinas 2021).

Terminas *natūralioji kalba* yra pažodinis vertimas iš anglų *natural language*. Tačiau jis ne visai tiksliai apibrėžia kalbas, kuriomis žmonės bendrauja tarpusavyje. Ta pati kalba (lietuvių ar kita) gali būti ir natūrali, ir nenatūrali. Aprašant klaidų atsiradimo priežastis automatinio vertimo metu, kaip viena jų paminima skirtinga žodžių tvarka sakinyje ir pateikiamas pavyzdys *He enjoys doing it – Jis mėgsta daryti tai*. Įvertinimas: „Lietuviškai **natūraliau** (paryškinta D. Š.): *Jis mėgsta tai daryti*“ (Rimkutė, Kovalevskaitė 2008a: 7–8). Vadinasi, automatinio vertimo metu pateiktas sakinys skamba nenatūraliai, tai – nenatūrali lietuvių kalba. Lygiai taip pat šia reikšme ir anglų mokslininkai vartoja žodį *natural*. Filas Kingas (Phil King) 2015 m., duodamas keletą patarimų, kaip versti iš anglų kalbos į polisintetines kalbas, rašė: „Keli anglų kalbos žodžiai gali būti išversti į jūsų kalbą vienu žodžiu su keliomis morfemomis, pvz., *He saw them* anglų kalboje yra trys žodžiai, bet kalo kalboje – tai vienas žodis su keturiomis morfemomis *egitarato*. Jeigu jūs kiekvieną anglų kalbos žodį versite atskiru jūsų kalbos žodžiu, **tai skambės labai nenatūraliai** (paryškinta D. Š.)“[2] (King 2015, 2 interneto nuoroda[3]), taigi, toks vertimas bus nenatūrali kalo kalba.

Vokiečių kalbos terminas *Nationalsprachen*, t. y. nacionalinės, tautų kalbos yra tikslesnis, todėl jis ir vartojamas šioje mokslo studijoje.

1.2. Truputis istorijos

Iki XX a. antrosios pusės visos gramatikos ir žodynai buvo spausdinami ant popieriaus ir skirti naudotis tik žmogui. Reikia pasakyti, kad ir „visa, daugiau nei 2 000 metų istoriją turinti kalbotyra, skirta žmogui“ (Marcinkevičienė 2002: 3). Prieš atsirandant kompiuteriams, kalba apskritai buvo laikoma vien tik žmonėms būdinga savybe. Kitų gyvūnų naudojamos signalinės sistemos buvo per daug paprastos ir primityvios, kad galėtų būti pavadintos „kalbomis“, o mechaniniai ir elektriniai prietaisai tegalėjo išsaugoti ir perduoti kodų sekas, kurias šifruodavo ir suprasdavo tik

[2] “Several words in English may be translated by one word in your language, with several morphemes in it. For example, ‘he saw them’ is three words in English, but one word in Kalo with four morphemes: egitarato. If you translate every word in English with a separate word in your language, **it may sound very unnatural** (paryškinta D. Š.)” (King 2015, 2 interneto nuoroda, 9 min. 25 sek.).

[3] Prieiga internete: https://www.youtube.com/watch?v=8ypQq5MvT24 [žiūrėta 2022-11-22].

žmogus (Winograd 1983: 23). 1942 m. Harvardo universitete (Harvard University) buvo sukurtas pirmasis pasaulyje kompiuteris MARK I (Schwanke 1991: 69). Gramatikos, kaip ir žodynai, persikėlė į kompiuterinę terpę ir atsirado naujas jų vartotojas – kompiuteris. Taip įvyko todėl, kad labai greitai tapo aišku, jog kompiuteriai gali apdoroti ne tiktai skaičius, bet ir kitus simbolius, pvz., raides, taigi, jie gali apdoroti ir kalbas (Winograd 1983: 23). Dirbtinio intelekto sferoje tautų kalbų apdorojimas tapo viena pagrindinių pritaikymo sričių (Chabris 1989: 71). Nuo 1968 iki 1978 m. tautų kalbų gramatinė analizė buvo pagrindinė dirbtinio intelekto tyrinėtojų darbo tema (Nirenburg 1987: 26). Tačiau „tautų kalbos, lengvai suprantamos žmonėms, pasirodė sunkiai įkandamos kompiuteriams“[4] (Jensen, Heidorn, Richardson 1993: 2). Todėl labai gerų rezultatų automatinio kalbų apdorojimo srityje iki šiol dar nepasiekta.

Pirmoji kompiuterių pritaikymo sritis, sulaukusi daug dėmesio, buvo tekstų vertimas iš vienos kalbos į kitą. 1947 m. Vorenas Vyveris (Warren Weaver) pirmasis iškėlė mintį apie kompiuterių panaudojimą vertimo darbams: „Įdomu, ar nebūtų įmanoma sukurti kompiuterio, kuris verstų“[5] (Hutchins, Sommers 1992: 25). 1948 m. Alanas Tiuringas (Alan Turing), dirbtinio intelekto pradininkas, išvardydamas būdus, kuriais gali pasireikšti kompiuterių „protas“, trečiuoju pamini kalbų vertimą (Hutchins, Sommers 1992: 27), nors automatinio vertimo ištakų galima būtų įžvelgti ir daug anksčiau. Pirmą kartą mechaninio vertimo idėjos buvo užfiksuotos XVII a. 1629 m. Renė Dekartas (René Descartes) siūlė rašyti knygas, sudarytas iš šifrų, o žodynuose visų kalbų atitikmenims turėjo būti suteiktas tas pats kodo numeris. 1661 m. pasirodžiusiame Johano Joachimo Becherio (Johan Joachim Becher) žodyne dešimčiai tūkstančių lotyniškų žodžių buvo suteikti kodai. Įdomu tai, kad ši knyga pavadinimu *Apie mechaninį kalbų vertimą: bandymas programuoti 1661 metais*[6] buvo perspausdinta visai neseniai, praėjus 300 metų nuo jos pasirodymo, t. y. 1962 m. (Hutchins, Sommers 1992: 21). Tačiau surasti ekvivalentus graikų, hebrajų, vokiečių, prancūzų, slavų ir arabų kalbomis, kaip buvo numatęs autorius, pasirodė ne taip paprasta. Vėliau lingvistai suprato, kad kalbų skirtumai yra tokie dideli, jog jų negali apimti vien tik

[4] “Natural language is easy for people and hard for machines” (Jensen, Heidorn, Richardson 1993: 2).

[5] “I have wondered if it were unthinkable to design a computer, which would translate” (Hutchins, Sommers 1992: 25).

[6] „Zur mechanischen Sprachübersetzung: ein Programmierungsversuch aus dem Jahre 1661 – Becher, 1962“ (Hutchins, Sommers 1992: 21).

žodynai, kad ir kaip „logiškai“ jie būtų sudaryti. Tai liudija 1903 m. spaudoje pasirodžiusi Vilhelmo Rygerio (Wilhelm Rieger) skaitmenimis koduota gramatika *Skaitmeninė gramatika, kuri kartu su žodynais leidžia mechaniškai versti iš vienos kalbos į visas kitas*[7]. Joje skaičiais koduojamos ne tik morfologinės kategorijos (linksnis, giminė, skaičius, laikas, asmuo ir pan.), bet ir sintaksinės, pvz., veiksmažodžių tranzityvumas (Rieger 1903: 107).

Praeito amžiaus pabaigoje kompiuterinės lingvistikos specialistai teigė, kad detaliai aprašyti automatinio vertimos sistemas nėra galimybės, nes visi duomenys dažniausiai yra įslaptinami. Kaip pavyzdį galima pateikti 1972 m., Vietnamo karo metu, sukurtą sistemą *Logos*, skirtą karinės tematikos tekstams. Ji atlikdavo vertimus iš anglų kalbos į vietnamiečių kalbą ir atvikščiai. 1982 m. į šią sistemą buvo įtraukta vokiečių kalba. Vokiečių–anglų kalbų versiją parengė firma *Siemens*, tačiau net vokiečių bendradarbiams baziniai žodynai ir duomenys buvo neprieinami (Schwanke 1991: 99). Tokia pati situacija būdinga ir mūsų Vytauto Didžiojo universiteto (toliau – VDU) anglų–lietuvių kalbų automatinio vertimo sistemai. Ją kūrė Sankt Peterburge, anglų–rusų kalbų vertimo sistemą pritaikant anglų–lietuvių vertimui. Iš kauniečių mokslininkų buvo pareikalauta tik tam tikrų lingvistinių duomenų apie lietuvių kalbą. Užsienio apžvalgininkai kalbinį teksto apdorojimą automatinio vertimo sistemose kartais vadina „juodąja dėže“: „[...] kalbinė sistemos dalis pateikiama kaip juodoji dėžė“[8] (Ananiadou 1987: 178). Taip paprastai vadinama bet kokia sistema, kurios vidinis veikimas yra paslėptas nuo vartotojo arba sunkiai suprantamas.

Padėtis iš esmės nepasikeitė iki pat šių dienų: „juodosios dėžės“ idėja tebeegzistuoja, tik jau dėl kitos priežasties. Anksčiau informaciją įslaptindavo žmonės, norėdami apsisaugoti nuo konkurentų, o šiuo metu ją „įslaptina“ patys kompiuteriai, nes dabar naudojamos iš esmės kitokios technologijos, kurios ir sukuria tą „juodąją dėžę“, tiksliau pasakius, kompiuteriai „paslepia“ savo darbo eigą nuo žmogaus (plačiau apie tai žr. 1.3.2 poskyryje).

[7] „Zifferngrammatik, welche mit Hilfe der Wörterbücher ein mechanisches Übersetzen aus einer Sprache in alle anderen ermöglicht“ (Rieger 1903).

[8] “[...] linguistic part of the system is supplied in the form of a “black box” (Ananiadou 1987: 178).

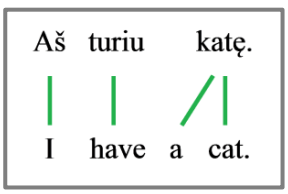
1.3. Šiuolaikiniai kalbos apdorojimo metodai

Tikriausiai nė vienas neginčys teiginio, kad kompiuteriai geriau už žmogų atlieka aritmetinius veiksmus. Jie puikiai susidoroja su formaliai aprašytomis užduotimis ir niekada nepadaro klaidų. Bet, pabandžius kompiuteriui pateikti tautų kalbų apdorojimo ar kitos intelektinės srities darbų, paaiškėjo, kad tokias užduotis labai sunku aprašyti formaliai. Todėl imta naudoti statistinius metodus, o vėliau buvo sukurti ir neuroniniai tinklai. Tačiau čia iškilo kita problema: priartinus kompiuterį prie žmogiškojo proto galimybių – imitavus neuronų veikimą – perimtas ir kitas žmogui būdingas faktas – klaidos („klysti žmogiška“).

1.3.1. Statistiniai metodai

Vorenas Vyveris, pirmasis iškėlęs mintį apie tai, kad kompiuteriai galėtų atlikti tautų kalbų vertimus, 1949 m. siūlė automatinio vertimo problemas spręsti pasitelkiant statistiką. Tačiau to meto mokslininkai greitai atsisakė tokių bandymų, matyt, dėl nedidelio tuometinių kompiuterių pajėgumo ir dėl nepakankamo kiekio tekstų, sukauptų elektronine forma. Po poros dešimtmečių, kai buvo surinkti tekstynai ir, naudojant statistinius metodus, buvo gauti geri sakytinės kalbos atpažinimo rezultatai, vėl sugrįžtama prie tikimybių teorija paremto automatinio vertimo (Brown ir kt. 1990: 79). Kanadoje tam buvo itin palankios sąlygos, nes čia parlamento medžiaga saugoma ir anglų, ir prancūzų kalbomis (dvi valstybinės kalbos), todėl greičiausiai susikaupė pakankamas lygiagrečių tekstų kiekis (Al-Onaizan ir kt. 1999: 1).

Taikant statistinius metodus, galima versti nesinaudojant nei gramatikos taisyklėmis, nei žodynu (Daudaravičius 2006: 13). Žodžių atitikimas tarp skirtingų kalbų vertimo metu nustatomas iš dvikalbių lygiagrečiųjų tekstynų (Nießen, Ney 2000: 1081). Labai paprastas lygiagrečiojo teksto pavyzdys pateikiamas 1 pav. Naujausios technologijos gali surasti atitikmenis ne tik tarp žodžių, bet ir tarp frazių, t. y. jos gali susieti, pvz., *pila kaip iš kibiro* su *raining cats and dogs*.



1 pav. Lygiagretusis lietuvių ir anglų kalbų sakinys

Naudodami lygiagrečiuosius tekstynus, kompiuteriai verčia tekstus iš vienos kalbos į kitą, tačiau pateikia tik apytikrį rezultatą[9] (Geitgey 2016). 2 pav. parodyta, kaip kompiuteryje atliekamas statistinis vertimas (Ranta 2017: 16). Pirmose dviejose eilutėse pateikta po du lygiagrečiuosius sakinius iš kiekvienos kalbos tekstyno. Išverstos atskiros sakinio atkarpos pažymėtos skirtingomis spalvomis. Trečioje eilutėje užrašytas verčiamas anglų kalbos sakinys. Išverstame švedų kalbos sakinyje *Denna stora hus är gul* (trečia eilutė) matyti, kad švedų kalbos žodį *namas* – *hus* pažymi tos pačios giminės žodis *denna*, kaip ir žodį *automobilis* – *bil* (pirma eilutė), tačiau tekstyno sakiniuose jų giminė skirtinga[10] – žodį *hus* pažymi žodis *detta* (antra eilutė). Tokios klaidos priežastis – giminės nebuvimas anglų kalboje.

2 pav. Vertimo, naudojant lygiagrečiuosius tekstynus, pavyzdys (Ranta 2017: 16)

Verčiant statistiniais metodais, iš daugelio tekstynuose sukauptų vertimo variantų ieškoma labiausiai tikėtino. Tačiau labiausiai tikėtinas variantas ne visada būna teisingas, t. y. ne visada būna tas, kuris pavartotas nagrinėjamame tekste.

Daug problemų iškyla tada, kai verčiami vadinamojo didelio kaitomumo kalbų tekstai (o būtent tokia ir yra lietuvių kalba), nes tekstynuose sunku apimti visas galimas kiekvieno žodžio formas. Latvių kalbininkai pabrėžia, kad metodai, gerai tinkantys mažai kaitomų kalbų analizei atlikti, didelio kaitomumo kalboms nėra efektyvūs, nes net ir labai didelės apimties tekstynuose gali nebūti rečiau pasitaikančių formų (Paikens, Rituma, Pretkalniņa 2013: 272). Statistinio vertimo esmė ta, kad jis nebando generuoti vieno tikslaus vertimo, bet sudaro daug galimų vertimo variantų ir

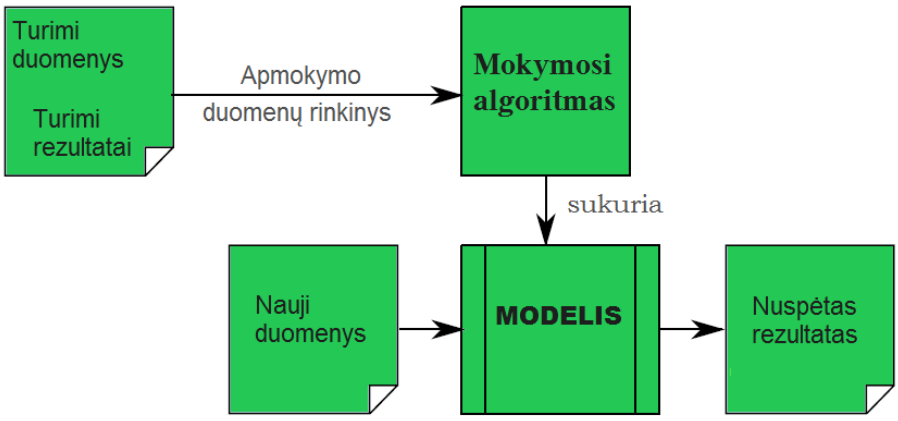
[9] "[...] computers can use parallel corpora to guess how to convert text from one language to another" (Geitgey 2016).

[10] Švedų kalboje yra išskiriamos dvi daiktavardžių giminės: bendroji (*utrum*) ir niekatroji (*neutrum*). Bendrosios giminės žodžiai turi artikelį *en*, o niekatrosios giminės – *ett* (3 interneto nuoroda: https://lt.wikipedia.org/wiki/%C5%A0ved%C5%B3_kalba). Žodis *automobilis* – *bil* yra bendrosios giminės, todėl jo artikelis yra *en* ir su juo vartojamas parodomasis įvardis *denna*, žodis *namas* – *hus* yra niekatrosios giminės: jo artikelis yra *ett*, o parodomasis įvardis turi būti *detta*. Kaip matyti iš 2 pav., žodžiui *namas* – *hus* neteisingai parinkta parodomojo įvardžio giminė (*denna*). Teisingas vertimas turėtų būti *Detta stora hus är gul*.

surikiuoja juos pagal tai, kiek tikėtina, kad kiekvienas jų yra teisingas. Apie teisingumo lygį sprendžiama tikrinant, kiek gautas vertimas yra panašus į mokymo duomenyse buvusius pavyzdžius (Geitgey 2016).

1.3.2. Neuroniniai tinklai

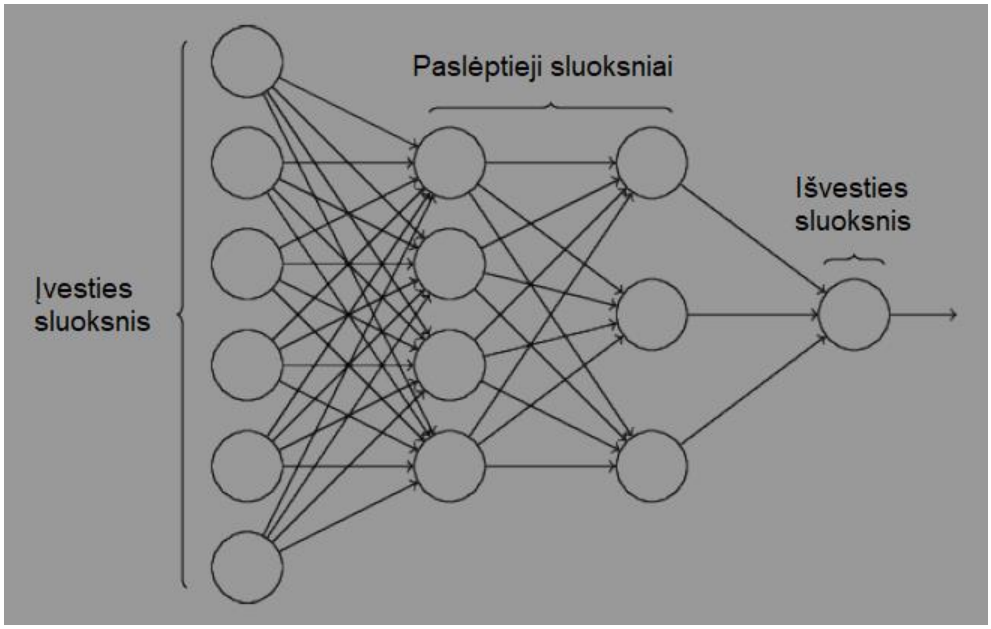
Pastaruoju metu statistinį vertimą išstumia kita vertimo metodika – neuroniniai tinklai. Šio metodo pagrindas – automatinis mokymasis (angl. *machine learning*). Naudojant tradicinį programavimą, kompiuteriui pateikiami pradiniai duomenys ir labai tiksliai nurodoma, kokias operacijas jis turi su jais atlikti, kad būtų gautas norimas rezultatas. Neuroninių tinklų atveju nepasakoma, kaip turi būti sprendžiamas uždavinys; mokymo metu pateikiami pradiniai duomenys ir rezultatas, kuris turi būti gautas, o kompiuteris pats sudaro taisykles[11] ir jas vėliau naudoja kitiems, naujiems, duomenims apdoroti (Nielsen 2018: 5). Kitais žodžiais tariant, kai naudojami tradicinio programavimo (taisyklėmis pagrįsti) metodai, geresnių rezultatų pasiekiama žmogui tobulinat programinės įrangos kodą, o, taikant automatinio mokymosi metodą, programos parašomos taip, kad jos pačios pakoreguoja savo atliekamą darbą naudodamos vis daugiau gaunamų duomenų. Tačiau joms ne visada pavyksta sėkmingai save patobulinti. Automatinio mokymosi struktūrinė schema pateikta 3 pav. (Berral ir kt. 2010: 4).



3 pav. Automatinio mokymosi struktūrinė schema (parengta pagal Berral ir kt. 2010: 4)

[11] "In the conventional approach to programming, we tell the computer what to do, breaking big problems up into many small, precisely defined tasks that the computer can easily perform. By contrast, in a neural network we do not tell the computer how to solve our problem. Instead, it learns from observational data, figuring out its own solution to the problem at hand" (Nielsen 2018: 5).

Kuriant neuroninius tinklus pagrindinis vienetas yra dirbtinis neuronas (angl. *artificial neuron*), t. y. matematinė funkcija, kuri yra suvokiama kaip biologinio neurono modelis (Žaliauskas 2017: 6). Dirbtiniai neuroniniai tinklai yra sudaryti iš tarpusavyje sujungtų dirbtinių neuronų. Gilusis neuroninis tinklas (angl. *deep neural network*) reiškia, kad jame yra keli paslėptieji neuronų sluoksniai (4 pav.). Pirmasis iš kairės sluoksnis vadinamas įvesties sluoksniu ir jame yra įvesties neuronai, pirmasis iš dešinės – išvesties sluoksniu ir jis apima išvesties neuronus, kurių gali būti vienas ar daugiau. Tarp jų esantys sluoksniai vadinami paslėptaisiais sluoksniais (Nielsen 2018: 19), kurių gali būti ne tik du, kaip parodyta šiame paveikslėlyje, bet ir labai labai daug.



4 pav. Gilusis neuroninis tinklas (parengta pagal Nielsen 2018: 19)

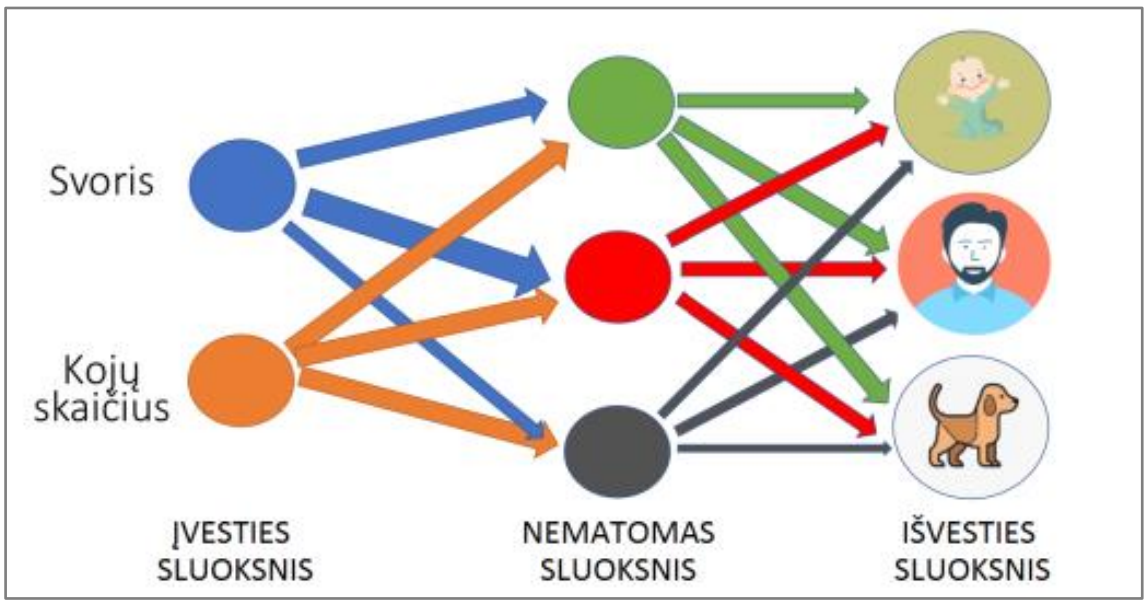
„Jei neuroninis tinklas yra visai mažytis, jį suprasti galima. [...] Bet labai didelis ir turintis tūkstančius neuronų sluoksnyje ir šimtus sluoksnių, tinklas tampa visai nesuprantamas. [...] Neįmanoma tiesiog žvilgtelėti į giliojo neurotinklo vidų ir pažiūrėti, kaip jis veikia“ (4 interneto nuoroda[12]). Pagal pateiktą programos kodą ir pateiktus apmokymo duomenis (įvestį ir rezultatą) kompiuteris susikuria modelį – nustato neuroninio tinklo koeficientų (vadinamųjų svorių) reikšmes. Tų koeficientų paprastai būna labai daug, o jų prasmė neaiški – neaišku, kaip tam tikrą koeficientą ar koeficientus neuroniniame tinkle reikia pakeisti, kad jis veiktų geriau. Todėl „[...] jau

[12] Prieiga internete: http://www.technologijos.lt/n/technologijos/it/S-61326/straipsnis/Tamsioji-dirbtinio-intelekto-paslaptis-niekas-is-tiesu-nesupranta-jo-veikimo [žiūrėta 2022-11-22].

dėl pačios savo prigimties gilusis mokymasis (angl. *deep learning*) yra itin tamsi juodoji dėžė“ (4 interneto nuoroda). Galima paminėti vieną sudėtingiausių neuroninių tinklų sistemų GPT-3, sukurtą 2020 m. Ji turi 69 sluoksnius, kurių kiekviename yra po kelias dešimtis tūkstančių neuronų, todėl susidaro apie 175 milijardus koeficientų (svorių). Šiai sistemai buvo naudojami superkompiuteriai, todėl apmokymui prireikė kelių mėnesių, su paprastais asmeniniais kompiuteriais tai būtų užtrukę daugiau nei 350 metų (Perez-Ortiz, Forcada, Sanchez-Martinez 2022: 149).

Buvo pastebėta, kad vertimo sistemos, kurios remiasi neuroniniais tinklais, gali „versti ir tarp tokių kalbų porų, kurioms tie tinklai nebuvo tiesiogiai apmokyti. Jeigu [...] sistema buvo apmokyta versti iš kalbos A į kalbą B, taip pat iš kalbos A į kalbą C, tai sistema sugeba versti ir iš kalbos B į C [...]. Tai reiškia, kad sistemos neuroninis tinklas apmokymo metu susikuria savo interlingvą, kuri suprantama tik jam pačiam“ (Ralys 2017: 13). Tačiau tokio vertimo kokybė gali būti šiek tiek prastesnė[13] (Johnson 2017: 350)

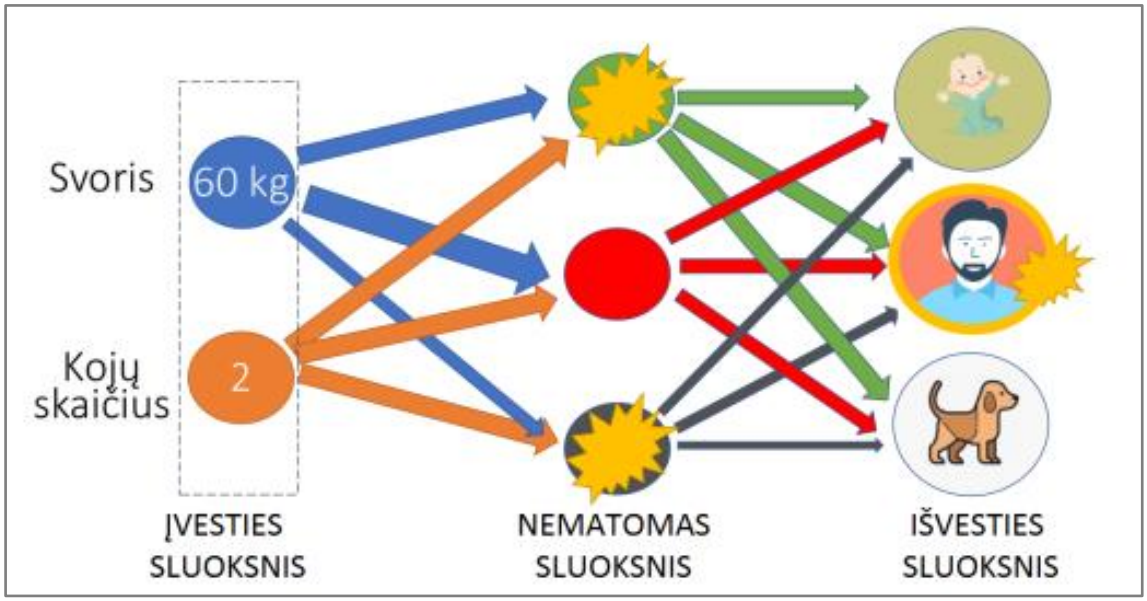
Populiariai parodyti neuroniniais tinklais pagrįstų sistemų veikimą galima pasitelkus nedidelį pavyzdį. 5 pav. pateikta tinklo, turinčio du neuronus įvesties sluoksnyje ir tris neuronus išvesties sluoksnyje, konfigūracija. Pagal du parametrus – kojų skaičių ir svorį – šis neuroninis tinklas gali atpažinti tris objektus: vaiką, vyrą ir šunį.



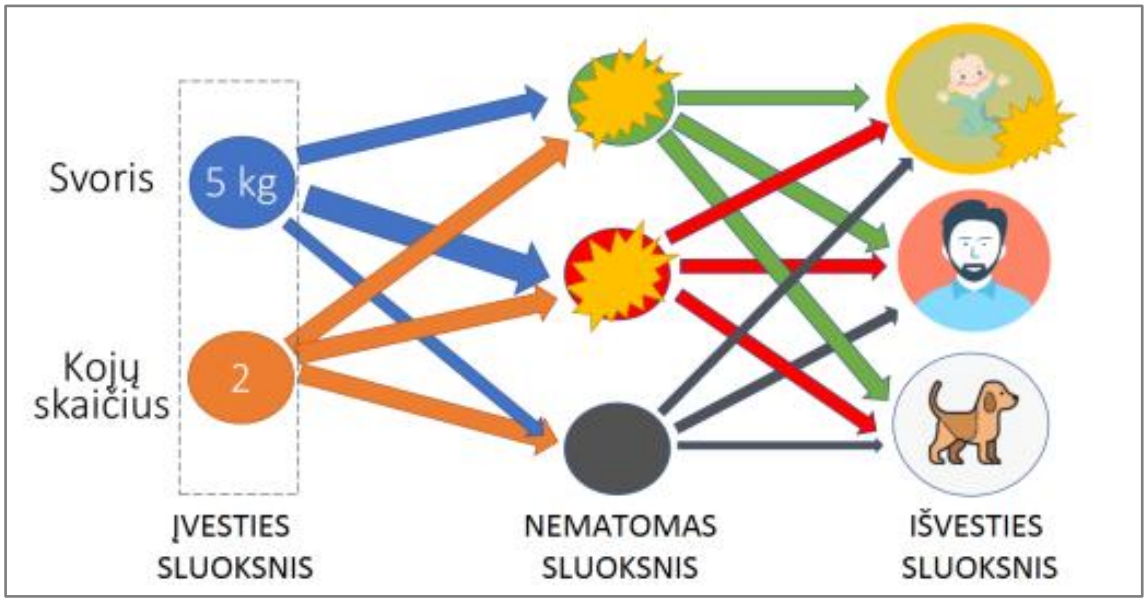
5 pav. Neuroninio tinklo konfigūracija (parengta pagal Liubinas 2021: 34)

[13] “...slightly lower translation quality” (Johnson ir kt. 2017: 350)

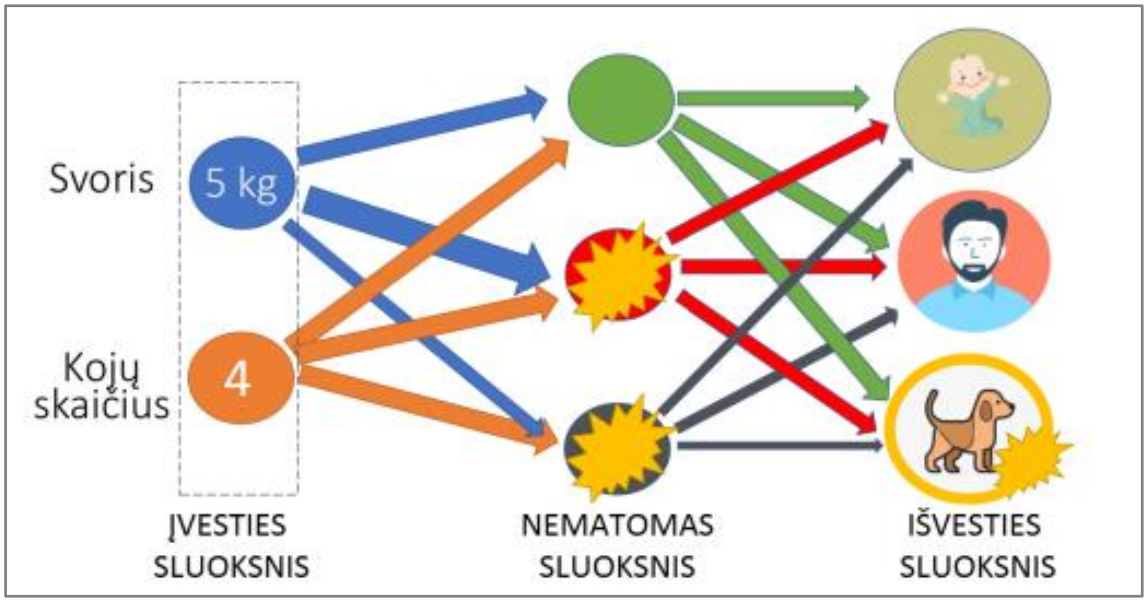
6 pav. pateiktas šio neuroninio tinklo veikimas atpažįstant vyrą (Liubinas 2021: 33). Jei įvesties sluoksnio neuronams pateikiami duomenys – svoris 60 kg ir kojų skaičius 2, – tuomet tinklo sprendimas yra, kad tai – vyras. Jei bus pateikti kiti duomenys, pvz., svoris 5 kg, o kojų skaičius liks tas pats – du, tinklo sprendimas pasikeis ir šiuo atveju jis bus: vaikas (7 pav.). 8 pav. pavaizduotas tinklo veikimas, pateikus duomenų rinkinį 5 kg ir 4 kojos.



6 pav. Neuroninio tinklo veikimas atpažįstant vyrą (parengta pagal Liubinas 2021: 33)



7 pav. Neuroninio tinklo veikimas atpažįstant vaiką (parengta pagal Liubinas 2021: 32)



8 pav. Neuroninio tinklo veikimas atpažįstant šunį (parengta pagal Liubinas 2021: 34)

Statistinių metodų ir neuroninių tinklų palyginimas. Abiejų metodų pagrindas yra automatinis mokymasis. Statistiniai metodai modelius sukuria naudodami n-gramas, t. y. kelių (n) žodžių junginius, tačiau vertimo metu tame pačiame sakinyje tos n-gramos traktuojamos kaip nepriklausomos viena nuo kitos. Neuroninių tinklų metodu modelis sudaromas atsižvelgiant į visą sakinį, o ne tik į n-gramas, todėl geriau susitvarkoma su derinimo klausimais. Tačiau apsiribojimas sakinio apimtimi sukelia daug problemų, kurios gali būti pastebėtos tiktai verčiant tekstą, o ne pavienius sakinius, pvz., negalima nustatyti įvardžio *it* atitikmens prieš tai buvusiuose sakiniuose (Kenny 2022: 37, 43).

Statistinio automatinio vertimo sistemose apmokymo metu sukuriami du modeliai: *vertimo* ir *derinimo*. Reikia pabrėžti, kad abu jie sudaromi visiškai be žmogaus pagalbos. Pirmajam modeliui paruošiama dvikalbė n-gramų tikimybių lentelė. Tokios lentelės pavyzdys italų–anglų kalboms parodytas 9 pav.

	English	Probability
a me piace	I like	0.78
a me piace	I should like to	0.11
a me piace	I admire	0.11

9 pav. Italų–anglų kalbų n-gramų lentelės fragmentas (Kenny 2022: 36)

Antrasis modelis – derinimo – yra vienakalbis ir taip pat paremtas n-gramomis. Jame surašomos tikimybės, koks žodis yra labiausiai tikėtinas po prieš tai buvusių žodžių, pvz., *Europarl* tekstyno duomenimis žodis *gorgonzola* (mėlynojo pelėsio sūris) po žodžių *I like* gali būti su tikimybe 0,024, tai reiškia, kad žodžių junginys *I like gorgonzola* buvo apmokymo duomenyse (konkrečiai jis pasitaikė keturis kartus), tačiau yra daugybė kitų žodžių, kurie labiau tikėtini po žodžių junginio *I like*.

Vertimo metu sudaroma daugybė galimų sakinio atitikmenų kitoje kalboje ir apskaičiuojama, kuris variantas yra labiausiai tikėtinas. Tam tikra labai nedidelė tikimybės dalis suteikiama ir tiems žodžiams, kurių apmokymo duomenyse nebuvo (Kenny 2022: 36–37).

Neuroniniai tinklai skiriasi nuo statistinio vertimo tuo, kaip sudaromi modeliai, nors abiem atvejais naudojamas automatinis mokymasis. Neuroniniai tinklai imituoja žmogaus neuronų veikimą, todėl duomenys modelyje vaizduojami visai kitu būdu nei statistiniuose metoduose. Vaizdavimo būdas paprastai pasirenkamas toks, kuris labiausiai atitinka poreikius, pvz., galima parašyti žodį *obuolys* ir galima nupiešti paveikslėlį, kuriame būtų obuolys, abiem atvejais bus pavaizduotas tam tikros rūšies vaisius, tačiau kompiuterinėmis priemonėmis galima patikrinti žodžio *obuolys* rašybą, o su paveikslėliu to padaryti negalima. Taigi, neuroniniams tinklams tinkamiausias žodžio pavaizdavimas yra vektorius, t. y. skaičių seka, pvz., žodis *obuolys* gali būti pavaizduotas vektoriumi [1.20, 2.80, 6.10], žodžio *kriaušė* vektorius tada galėtų būti [1.20, 2.80, 5.50] – kodai yra panašūs, nes ir objektai yra artimos reikšmės (Kenny 2022: 41–42). Šioje vietoje modeliuojama žmogaus galimybė apibendrinti, t. y. tai, ką mes išmokome, galime panaudoti ateityje, panašiose, bet ne tokiose pačiose situacijose, pvz., žmogus, neturi iš naujo mokytis vairuoti, jei sėda prie kito automobilio vairo, ar važiuoja kitomis gatvėmis nei mokymosi metu (Perez-Ortiz, Forcada, Sanchez-Martinez 2022: 147).

1.3.3. Rezultatų tikslumas ir patikimumas

Naudojant neuroninius tinklus įvairioms atpažinimo užduotims atlikti buvo pasiekta gana gerų rezultatų. Bene labiausiai jų gebėjimai pasireiškė atpažįstant vaizdus ir sakytinę kalbą (Szegedy ir kt. 2014: 1).

1.3.3.1. Vaizdų atpažinimas

Vaizdams atpažinti naudojama speciali neuroninio tinklo rūšis – sąsūkinis neuroninis tinklas (angl. *convolutional neural network*). „Pagrindinė konvoliucinių neuroninių tinklų struktūra ir veikimas buvo įkvėpti smegenyse vykstančių vaizdo atpažinimo procesų“ (5 interneto nuoroda[14]), tačiau techninis šio metodo įgyvendinimas yra iš esmės kitoks (Fukushima 1980: 193). Kompiuteriai paveikslėlius mato kitaip negu žmogaus akis. Kompiuterių pasaulis susidaro vien iš skaičių, todėl visi vaizdai koduojami skaičių rinkiniais dvimatėje erdvėje – plokštumoje (Gulbinas 2019: 13). 10 pav. parodyta, kaip paveikslėlį mato žmogaus akys ir kaip jis vaizduojamas kompiuterio vidiniame formate.

10 pav. Vaizdas, matomas žmogaus akimis ir kompiuterio (Gulbinas 2019: 13)

Šie skaičiai, nors ir beprasmiški žmogui, tačiau kompiuteriui yra vieninteliai prieinami duomenys atpažįstant vaizdus. Atpažinimo metodas yra toks: sąsūkiniai neuroniniai tinklai naudoja filtrus, kad surastų tam tikras vaizdo savybes, pvz., briaunas. Filtras slenka per vaizdą ir tikrina, ar yra funkcija, t. y. vaizdo savybė, kurią jis turi aptikti. Jei kurioje nors vaizdo dalyje yra filtro ieškoma savybė, sąsūkos operacija generuoja didelį skaičių, jei tokios savybės nėra, sugeneruotas skaičius bus mažas. Šie duomenys perduodami kitiems tinklo sluoksniams (Gulbinas 2019: 14).

Populiariai kompiuteriu atliekamo vaizdų atpažinimo užduotis parodyta 11 pav., t. y. kompiuteris turi nuspręsti, kas tai – keksiukas ar šuniukas (Wong 2017: 38). Tinklo sprendimas priklauso nuo neuronų, išsidėsčiusių dešimtyse ar net šimtuose glaudžiai susijusių sluoksnių, veikimo. Pirmojo sluoksnio neuronai gauna pradinius duomenis, pvz., vaizdo taško intensyvumą, ir, atlikę skaičiavimus, sukuria naują signalą. Kiekvieno žemesnio neuronų sluoksnio rezultatai perduodami aukštesnio lygio neuronams ir taip informacija pereina nuo įvesties iki išvesties, kol

[14] Prieiga internete: https://lt.wikipedia.org/wiki/Konvoliucinis_neuroninis_tinklas [žiūrėta 2022-11-22].

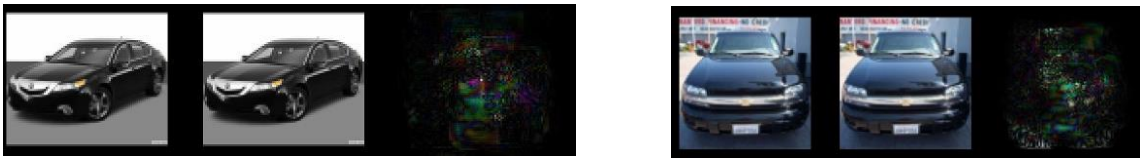
gaunamas rezultatas. Pavyzdžiui, „[...] sistemoje, skirtoje šunų atpažinimui, žemesni sluoksniai atpažįsta tokius paprastus dalykus, kaip siluetai ir spalvos; aukštesni sluoksniai atpažįsta sudėtingesnius, pavyzdžiui, kailį ar akis; aukščiausias sluoksnis visa tai identifikuoja kaip šunį“ (4 interneto nuoroda).

11 pav. Vaizdų atpažinimo užduotis (Wong 2017: 38)

Naujausi neuroninių tinklų pasiekimai yra tokie, kad jų atpažinimas beveik prilygsta žmogaus galimybėms, tad natūraliai iškilo klausimas, kuo skiriasi žmogaus ir neuroninių tinklų matymas (Nguyen, Yosinski, Clune 2015: 1).

Buvo atlikti du tyrimai, kurių metu analizuota, kuo skiriasi žmogaus ir neuroninių tinklų vaizdo atpažinimas. Pirmajame buvo parodyta, kad labai nedideli, žmogui nepastebimi vaizdo pakeitimai (trikdžiai) gali sukelti neuroninių tinklų klaidas (neatpažįstamas žmogui gerai žinomas objektas). Kito tyrimo metu tam tikrais evoliuciniais algoritmais buvo dirbtinai generuojami paveikslėliai, kurie žmogui yra visai nesuprantami, o neuroniniai tinklai juos laiko gerai pažįstamais daiktais su labai didele tikimybe.

Pirmojo tyrimo metu buvo analizuojamos atpažinimo klaidos, atsirandančios dėl trikdžių, kurie žmogui yra nepastebimi. Ir tai nėra susiję su mokymo problemomis, nes tas pats trikdis sukelia klaidingą atpažinimą įvairiems tinklams, kurie buvo apmokomi naudojant ne tuos pačius duomenų rinkinius. 12 pav. pateiktas automobilių atpažinimo pavyzdys (Szegedy ir kt. 2014: 6).

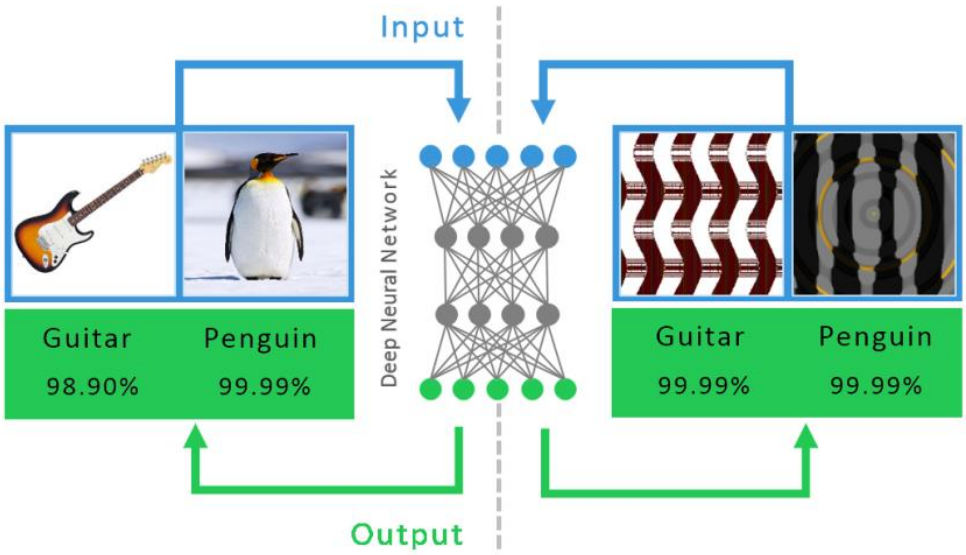


a) **b)**

12 pav. Automobilių atpažinimo klaidos (Szegedy ir kt. 2014: 6)

Abiejose 12 pav. dalyse – tiek a), tiek b) – kairėje pusėje esanti automobilio nuotrauka buvo atpažinta kaip automobilis, tačiau vidurinė nuotrauka abiem atvejais nebuvo priskirta automobiliams. Dešinieji paveikslėliai rodo skirtumus tarp kairiajame ir viduriniame paveikslėlyje esančių automobilių, kurių žmogus pačiose nuotraukose pastebėti negali.

Atliekant antrąjį tyrimą bandyta parodyti klaidų atsiradimą kita kryptimi. Paaiškėjo, kad labai nesunkiai galima sukurti paveikslėlius, kurie žmogui yra visai nesuprantami, bet neuroninis tinklas juos laiko gerai žinomais daiktais (Nguyen, Yosinski, Clune 2015: 428). 13 pav. parodyta, kaip neuroninis tinklas, gerai atpažįstantis gitarą ir pingviną, tiems patiems objektams priskiria žmogui visai nepanašius į juos vaizdus (tikimybė – 99,99 proc.).



13 pav. Klaidingas objektų priskyrimas gitarų ir pingvinų klasėms (parengta pagal Nguyen, Yosinski, Clune 2015: 428)

Žinoma, šiuo metu ieškoma būdų, kaip išvengti tokių situacijų, bet kol kas neuroniniai tinklai 100 proc. patikimų rezultatų dar neduoda.

Tinklalapis *Parts-of-speech.Info* kiekvienam vartotojo įvestam žodžiui nurodo kalbos dalį. Programinės įrangos pagrindą sudaro Stenfordo universiteto (Leland Stanford Junior University) morfologinis analizatorius *Stanford University Part-Of-Speech-Tagger*. Jo veikimas remiasi neuroniniais tinklais ir šalia teksto analizės pateikiama pastaba, kad gali būti, jog 100 proc. tikslumas niekada nebus pasiektas[15] (6 interneto nuoroda[16], skirtukas *about Part-of-speech.Info*). Vokiečių kalbos apmokymo duomenis šiam analizatoriui pateikė Zarlando universitetas (Universität des Saarlandes). Jie apima 50 morfologinių žymų. Tačiau paaiškinama, kad „automatinio mokymosi metodas veikia taip, kad, grubiai pasakius, programinė įranga šeriama tekstais, kuriuose žmogaus yra nurodyta, kokiai kalbos daliai koks žodis priklauso. Ir programinė įranga iš gautų duomenų turi pati susikurti taisykles (pvz., kad po artikelio su labai didele tikimybe eina būdvardis arba daiktavardis) ir privalo nurodyti kalbos dalį net ir tiems žodžiams, kurių ji niekada nėra „mačiusi“ [...]. Kompiuteriui tai yra labai sunki užduotis ir todėl nustatant kalbos dalį **daroma daug klaidų**“[17] (7 interneto nuoroda[18], skirtukas *über Wortarten.Info*).

Pagrindinė priežastis, dėl kurios visose neuroninius tinklus naudojančiose sistemose atsiranda klaidų, labai populiariai paaiškinta *Lietuvių kalbos išteklių informacinėje sistemoje – E. kalba* (8 interneto nuoroda[19]) diakritinių ženklų pavyzdžiu. Aprašyme apie tai, kaip atstatomi lietuvių kalbos diakritiniai ženklai atliekant interneto tekstų analizę, pasakyta, kad šiam tikslui pasiekti naudojamas kalbos modelis, kuriame yra iš didelių vienakalbių tekstynų išmoktos žodžių n-gramos (greta einančių žodžių sekos) ir tų n-gramų tikimybės. Tais atvejais, kai ženklų atstatymo variantai gali būti du, pvz., *sunelis susirgo* gali būti *sūnelis susirgo* ir *šunelis susirgo*, arba netgi trys, pvz.,

---

15 “Please be aware that these machine learning techniques might never reach 100 % accuracy” (6 interneto nuoroda, skirtukas *about Part-of-speech.Info*).

16 Prieiga internete: https://parts-of-speech.info/ [žiūrėta 2022-11-22].

17 „Das Verfahren mit den Trainingsdaten (maschinelles Lernen) funktioniert grob gesagt so, dass eine Software mit Texten gefüttert wird, bei denen einmal von Menschen hinterlegt wurde, welches Wort von welcher Wortart ist. Die Software kann dann daraus Regeln ableiten – z. B. dass nach einem Artikel mit hoher Wahrscheinlichkeit ein Adjektiv oder Nomen kommt – und so auch Wörtern Wortarten zuordnen, die sie vorher noch nie „gesehen“ hat. [...] ist es für den Rechner trotzdem eine ziemlich schwierige Aufgabe und bei der Erkennung **entstehen Fehler**“ (7 interneto nuoroda, skirtukas *über Wortarten.Info*).

18 Prieiga internete: https://wortarten.info/ [žiūrėta 2022-11-22].

19 Prieiga internete: https://pasauliolietuvis.lt/beribis-lietuviu-kalbos-pasaulis-skaitmeniniu-istekliu-sistemoje-e-kalba/ [žiūrėta 2022-11-22].

rastas, *rąstas* ir *raštas* (Kapočiūtė-Dzikienė, Davidsonas, Vidugirienė 2017: 509), visada bus imamas tik tas variantas, kuris apmokymo duomenyse pasitaikė daugiau kartų (9 interneto nuoroda[20]), t. y. jis yra labiau tikėtinas. Tačiau tai gali būti ir ne tas atvejis, kuris iš tikrųjų pavartotas tekste.

1.3.3.2. Automatinis vertimas

Neuroniniams tinklams bent jau kol kas geriau sekasi atpažinti objektus nuotraukose negu atlikti tautų kalbų apdorojimo užduotis (10 interneto nuoroda[21]).

Dar 2017 m. pirmosios anglų–lietuvių kalbų automatinio vertimo sistemos vadovas Vaidas Repečka teigė, kad „[...] šiandien geriausią vertimo kokybę užtikrina neuroniniais tinklais ir automatiniu mokymusi pagrįstos vertimo sistemos. Tačiau jų naudojimas problemiškas, nes taip iškraipoma fleksinių kalbų struktūra, klaidingai išverstos teksto dalys perkeliamos į kitus tekstus ir t. t.“ (11 interneto nuoroda[22]). Lietuvių kalbai tinkamiausias yra taisyklėmis pagrįstas automatinio vertimo metodas, „kai originalo kalbos tekstas „išnarstomas“ žodžio ir sakinio dalimis ir vėl „sudedamas“ kitoje kalboje“ (11 interneto nuoroda).

eTranslation. Viena naujausių neuroninių tinklų metodu veikiančių automatinio vertimo sistemų yra *eTranslation* (12 interneto nuoroda[23]). Ji skirta teisiniams tekstams versti. Gana išsami šios sistemos veikimo analizė buvo pateikta 2019 m. vykusiame antrajame ELRC[24] seminare Lietuvoje. Be įvardytų pagrindinių šios vertimo sistemos privalumų, kad *eTranslation* „dažnai pateikia gerus ar mažai taisytinus išversto teksto gabalus; beveik taisyklinga gramatika (tinkamos žodžių formos); atsižvelgia į kontekstą“, buvo nurodyti ir kol kas dar pastebimi trūkumai: „Prasmės iškraipymas (pvz., *active and healthy ageing* išversta *aktyvus ir sveikas*

[20] Prieiga internete: https://ekalba.lt/public#/sentimentAnalysis/about [žiūrėta 2020-04-17].

[21] Prieiga internete: https://www.aidas.lt/lt/mokslas-ir-it/article/22560-09-27-vdu-mokslininkai-vysto-dirbtinio-intelekto-technologiju-sprendimus-lietuviu-kalbai-kodai-bus-perduoti-visuomenei [žiūrėta 2022-11-22].

[22] Prieiga internete: http://alkas.lt/2017/11/03/kalbos-technologijos-butina-salyga-kalbai-gyvuoti/ [žiūrėta 2022-11-22].

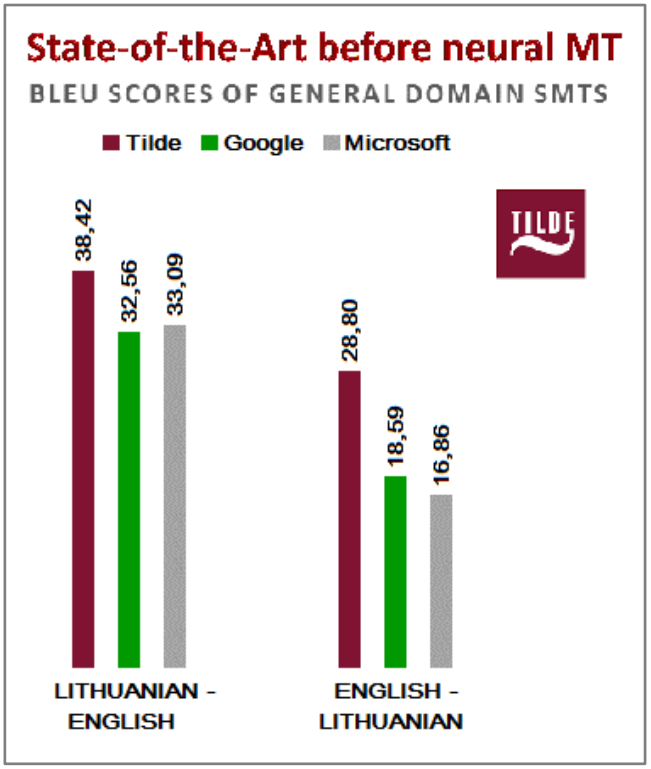
[23] Prieiga internete: https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en [žiūrėta 2022-11-22].

[24] *European Language Resource Coordination* – Europos kalbų išteklių koordinavimas.

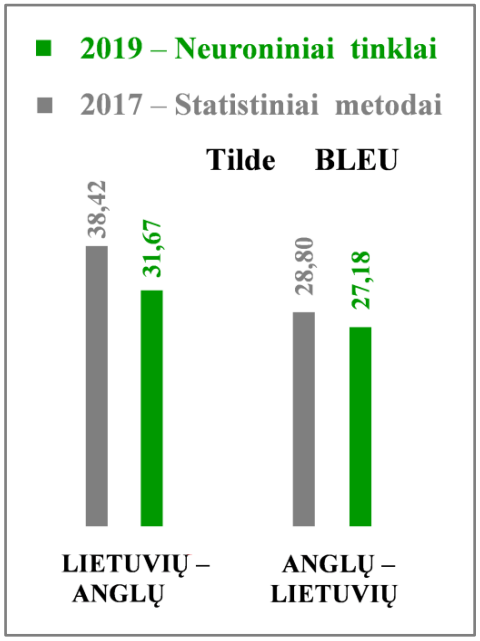
senėjimas; tokios realijos kaip *aktyvus senėjimas* nėra, žmogaus vertimas būtų: *vyresnio amžiaus žmonių aktyvumas ir sveikata*). Kiti minusai: terminijos nenuoseklumas; iš piršto laužti (pačios sistemos susikurti) žodžiai, terminai, gramatinės formos.“ Kaip problema buvo įvardyta, kad „gramatiškai sklandus vertimas gali slėpti svarbias prasmės klaidas“ (Zaikauskas 2019: 22–23). Pačios sistemos mokymasis buvo stebimas analizuojant gautų rezultatų pokyčius, kai versti buvo pateikiamas tas pats sakinys *Every single day somewhere in the world indigenous peoples are being dispossessed of their ancestral lands, territories and resources*. 2018 m. spalio mėn. buvo gautas toks vertimas: *Kiekvieną dieną pasaulio čiabuvių tautose yra* ***apykažiamos*** *jų protėviai, teritorijos ir ištekliai.* Žodis *apykažiamos* yra pačios vertimo sistemos susikurtas: ji analizuoja kontekstą ir pagal tam tikrą algoritmą (detalių veiksmų seką) parenka geriausią variantą. 2019 m. sausio mėn. verstas tas pats sakinys skambėjo šiek tiek geriau (bent jau nėra nesančių lietuvių kalboje žodžių): *Kiekvieną dieną visame pasaulyje čiabuvių tautos* ***nepaiso*** *savo protėvių žemių, teritorijų ir išteklių*; tačiau tikslaus vertimo čia dar nėra. Šis pavyzdys rodo, kad neuroninių tinklų metodu pagrįstos automatinio vertimo sistemos iš tikrųjų gali pačios kai ką išmokti, tačiau tas mokymasis ne visada būna sėkmingas, ir tai iliustruoja dar vienas pavyzdys. Tuo pačiu metu, 2018 m. spalio mėn., į lietuvių kalbą buvo verčiamas kitas anglų kalbos sakinys: *In point 1 of Annex I to Regulation (EC) No 1275/2008, the entry ‘Dish washing machines’ is deleted.* Jis buvo išverstas taip: *Reglamento (EB) Nr. 1275/2008 I priedo 1 punkte įrašas „****Diškių skalbimo mašinos****“ išbraukiamas.* Antrojo etapo metu, 2019 m. sausio mėn., gautas jau kitas vertimo variantas, tačiau ne geresnis už pirmąjį*: Reglamento (EB) Nr. 1275/2008 I priedo 1 punkte išbraukiamas įrašas „****Išk plovimo mašinos****“* (Zaikauskas 2019: 18).

Tilde. 2017 m. duomenimis, *Tilde* atliekami lietuvių kalbos vertimai į anglų kalbą ir iš jos buvo geresnės kokybės nei *Google* ar *Microsoft* (Skadiņš 2017: 22). Visose trijose vertimo sistemose buvo naudojami statistiniai metodai. 14 pav. pateikta jų analizė pagal BLEU (BiLingual Evaluation Understudy) įverčius, t. y. skaičiuojama, kokią dalį sakinių kompiuteris išvertė teisingai, palyginti su žmogaus vertimu.

2019 m. *Tilde* parengė automatinio vertimo sistemą, veikiančią neuroninių tinklų pagrindu. Jos rezultatai pagal BLEU įverčius yra 31,67 lietuvių–anglų kryptimi ir 27,18 anglų–lietuvių kryptimi (Pinnis, Krišlauks, Rikters 2019: 332), taigi, nors ir nedaug, tačiau nusileidžia statistinius metodus naudojančiam *Tilde* vertimui (15 pav.).

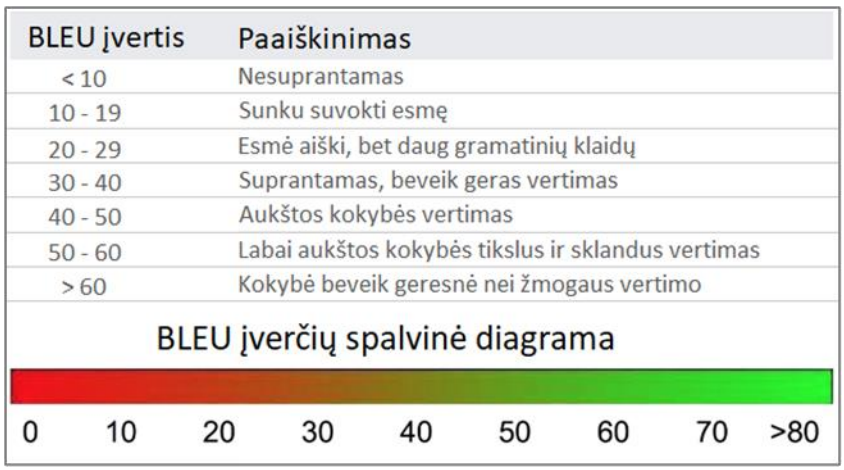


14 pav. Automatinio vertimo sistemų palyginimas pagal BLEU įverčius (parengta pagal Skadiņš 2017: 22)



15 pav. Neuroninių tinklų ir statistinių metodų palyginimas automatinio vertimo sistemoje *Tilde* pagal BLEU įverčius

Šioje vietoje vertėtų plačiau paaiškinti BLEU įverčių reikšmes. BLEU įvertis skaičiuojamas pagal matematines formules ir gali įgyti reikšmes nuo 0 iki 1, tačiau dažniausiai pateikiama jo procentinė išraiška (taip yra parodyta ir 14 pav. bei 15 pav.). Reikia atkreipti dėmesį, kad net žmogaus atliktas vertimas nepasiekia tobulo vertimo įverčio 1.0[25]. 16 pav. pateikiamas BLEU reikšmių aprašymas išversto teksto supratimo požiūriu (13 interneto nuoroda[26]).



16 pav. BLEU įverčių reikšmių paaiškinimas (parengta pagal 13 interneto nuorodą)

Kad žmogaus vertimas nepasiekia tobulo įverčio, galima teigti jau vien todėl, kad kartais sakinyje, kurį reikia išversti, nėra pakankamai informacijos, reikalingos išverstam sakiniui, t. y. išverstas sakinys kartais turi daugiau informacijos nei originalus. Prancūzų kalbos tekstas pasako daugiau negu anglų kalbos, pvz., kai sakinio veiksniu eina įvardis *aš*: prancūzų kalbos sakinyje šiuo atveju atsispindi veiksnio giminė, anglų kalbos sakinyje tokios informacijos nėra (Kenny 2022: 25).

1.4. Skyriaus išvados

Sieti kalbą su skaitmenimis žmonės pradėjo jau labai seniai. Dar viduramžiais gimė pirmosios idėjos automatizuoti vertimą iš vienos kalbos į kitą. Tačiau įgyvendintos jos buvo tik atsiradus kompiuteriams. Naujausios technologijos – neuroniniai tinklai, kuriuose modeliuojamas žmogaus neuronų darbas, – leido pasiekti gana gerų rezultatų vaizdų, sakytinės kalbos atpažinimo, automatinio vertimo srityse,

[25] "Note that even human translators do not achieve a perfect score of 1.0" (13 interneto nuoroda).

[26] Prieiga internete: https://cloud.google.com/translate/automl/docs/evaluate [žiūrėta 2022-11-22].

tačiau jų darbas dar neprilygsta žmogaus galimybėms. Bandymais buvo parodyta, kad neuroninius tinklus naudojančios sistemos yra gana greit pažeidžiamos, ir nedideli pakitimai (triukšmai) įvedamuose duomenyse gali sukelti daug ir esminių jų darbo klaidų. Neuroninių tinklų pagrindą sudaro automatinis mokymasis. Sistemos tobulina savo darbą gaudamos vis daugiau apmokymo duomenų, tačiau joms ne visada pavyksta save sėkmingai patobulinti, ir jau atvirai pripažįstama, kad gali būti, jog 100 proc. tikslumas niekada nebus pasiektas.

2. ANOTUOTI TEKSTYNAI

Tekstynas – tai didelis kiekis tekstų, sukauptų elektronine forma ir naudojamų daugiausia statistinės analizės tikslams (14 interneto nuoroda[27]). Esminis teksto ir tekstyno skirtumas yra tas, kad tekstyno niekas ištisai neskaito kaip teksto – jis analizuojamas tik su programinėmis priemonėmis (Marcinkevičienė 2000: 8).

Pirmasis lietuvių kalbos tekstynas, kurį sudaro 1 200 000 žodžių, buvo surinktas XX a. pabaigoje Matematikos ir informatikos institute (toliau – MII). Jis priklauso pirmosios kartos tekstynams, kurių apimtis – apie 1 mln. žodžių. Šis tekstynas buvo ruošiamas remiantis Prahos lingvistų naudota metodika: jį sudaro originalūs, neverstiniai keturių funkcinių kalbos stilių – publicistinio, kanceliarinio, beletristinio ir mokslinio – tekstai (Grumadienė 2002: 25), kurie buvo renkami 1990–1995 m. Surinkta po 300 000 kiekvieno funkcinio stiliaus žodžių, tiksliau, po 300 teksto atkarpų, atsitiktinai imant po 1 000 rišlaus teksto žodžių (Grumadienė, Žilinskienė 1997: VII). Tekstynas buvo morfologiškai anotuotas naudojant V. Zinkevičiaus morfologinės analizės programinę įrangą ir vėliau ranka buvo panaikintas daugiareikšmiškumas (Grumadienė 2002: 27). Šio tekstyno pagrindu išleisti du žodynai: 1997 m. *Dažninis dabartinės rašomosios lietuvių kalbos žodynas: mažėjančio dažnio tvarka* (Grumadienė, Žilinskienė 1997) ir 1998 m. *Dabartinės rašomosios lietuvių kalbos dažninis žodynas: abėcėlės tvarka* (Grumadienė, Žilinskienė 1998). Pats tekstynas viešai internete nėra prieinamas (Marcinkevičienė 2000: 16).

Vėliau, 2002 m., pasirodė kitas, jau viešai internete publikuojamas tekstynas (15 interneto nuoroda[28]), kuris buvo pradėtas rinkti 1992 m. Tai VDU *Dabartinės lietuvių kalbos tekstynas*. Jis buvo kuriamas pagal kitus atrankos principus – remiantis bendrinės kalbos samprata, kuri yra artimesnė amerikiečių standartinei kalbai (angl. *standard language*) ir apima ne tik rašytinę, bet ir buitinę bei kitą leksiką (Grumadienė 2002: 26). Tai subalansuotas tekstynas, „[...] nes iš turimo gautų tekstų archyvo į tekstyną pagal pasirinktas proporcijas perkeliama tik dalis tekstų“. 70 proc. sudaro periodika, 26 proc. – neperiodiniai leidiniai ir likę 4 proc. – verstinė literatūra. Jis apibrėžiamas kaip „didelis, neanotuotas ir nekoduotas, į skaitytoją orientuotas [...] rašytinės lietuvių kalbos tekstynas“ (Marcinkevičienė 1997: 77).

[27] Prieiga internete: https://en.wikipedia.org/wiki/Text_corpus [žiūrėta 2022-11-22].

[28] Prieiga internete: http://tekstynas.vdu.lt/tekstynas/menu?page=about [žiūrėta 2022-11-22].

2.1. Tekstynų rūšys

Pagal apimtį tekstynai skirstomi į mažus ir didelius arba, kitos klasifikacijos terminais įvardijus, į statinius ir dinaminius. Statiniai tekstynai apima mirusių kalbų, pvz., lotynų kalbos, rašytinius paminklus; mirusio rašytojo, pvz., Šekspyro, veikalus; Šventąjį Raštą ir kt. Šiems tekstynams būdinga tai, kad jie yra užbaigti ir niekada nebebus plečiami (16 interneto nuoroda[29]). Dinaminiai tekstynai pildomi nuolat (pvz., publicistikos), todėl gali būti labai dideli. Bene daugiausia naudojamas yra *Dabartinės Amerikos anglų kalbos tekstynas* COCA[30], apimantis daugiau kaip milijardą žodžių. Jis buvo rinktas nuo 1990 iki 2019 m. kasmet paruošiant po daugiau kaip 25 mln. žodžių iš aštuonių sričių: grožinės literatūros, populiarių žurnalų, laikraščių, akademinių tekstų ir kt. (17 interneto nuoroda[31]). Didžiausias yra internetinis tekstynas *iWeb*[32], turintis 14 milijardų žodžių (18 interneto nuoroda[33]).

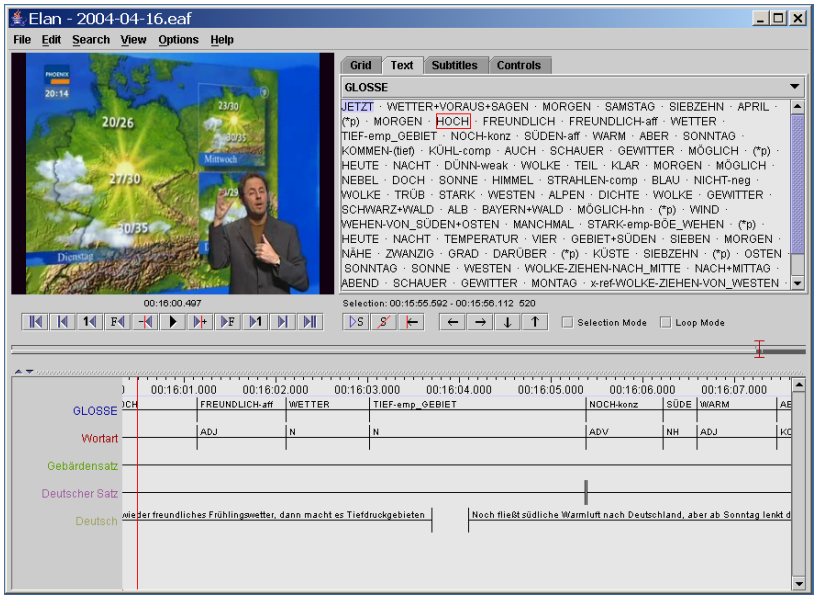
Pagal informacijos saugojimo formą gali būti teksto, garso įrašų, vaizdo įrašų (gestų kalba) tekstynai. Pirmasis Vokietijoje surinktas gestų kalbos tekstynas apėmė orų prognozės pranešimus (Bungeroth ir kt. 2006: 3). 17 pav. galima matyti ne tik žodžio vaizdą, bet ir morfologinę informaciją apie jį, pvz., kalbos dalį. 2020 m. pateikta trečioji DGS (vok. ***D****eutsche* ***G****ebärden****s****prache* – vokiečių gestų kalba) tekstyno versija (Jahn 2020).

2005 m. VDU buvo surinktas *Rišlaus teksto garsynas*, kurį sudaro apie 17,5 val. dviejų diktorių (vyro ir moters) šnekos įrašai. Juose yra iš viso 114 130 žodžių, 33 645 – skirtingi (19 interneto nuoroda[34]).

Pagal kalbų skaičių tekstynai gali būti vienakalbiai, dvikalbiai, daugiakalbiai ir kt. Europos Sąjungoje kaupiamas teisės aktų tekstynas *EUR-Lex*, apimantis visas ES kalbas (20 interneto nuoroda[35]). Vienu pirmųjų daugiakalbių tekstynų (tiesa, kol kas sudarytų tik iš vieno teksto) laikomas Rozetos akmuo (18 pav.), ant kurio buvo iškaltas tas pats tekstas – Egipto valdovo Ptolemajo V įsakymas, datuotas 196 m. pr. m. e., –

[29] Prieiga internete: https://de.wikipedia.org/wiki/Textkorpus [žiūrėta 2022-11-22].
[30] COCA – Corpus of Contemporary American English.
[31] Prieiga internete: https://www.english-corpora.org/coca/ [žiūrėta 2022-11-22].
[32] iWeb: The Intelligent Web-based Corpus.
[33] Prieiga internete: https://www.english-corpora.org/ [žiūrėta 2022-11-22].
[34] Prieiga internete: https://www.vdu.lt/cris/handle/20.500.12259/41313?mode=full [žiūrėta 2022-11-22].
[35] Prieiga internete: https://ec.europa.eu/social/main.jsp?catId=521&langId=lt [žiūrėta 2022-11-22].

užrašytas dviem kalbomis, bet trimis rašmenimis: Egipto hieroglifų raštu (viršutinė dalis, 14 eilučių), Egipto demotiškuoju raštu (vidurinė dalis, 32 eilutės) ir senąja graikų kalba (apatinė dalis, 54 eilutės) (21 interneto nuoroda[36]).



17 pav. Vokiečių gestų kalbos tekstyno pavyzdys (Bungeroth ir kt. 2006: 3)

18 pav. Rozetos akmuo, kuriame tas pats tekstas iškaltas dviem kalbomis (22 interneto nuoroda[37])

[36] Prieiga internete: https://lt.wikipedia.org/wiki/Rozetos_akmuo [žiūrėta 2022-11-22].

[37] Prieiga internete: https://istorijatau.lt/rubrikos/zodynas/rozetes-akmuo [žiūrėta 2022-11-22].

Sukaupus tekstynus, išsiskyrė nauja kalbotyros šaka – tekstynų lingvistika[38]. Vienas iš jos naudojamų tyrimo metodų vadinamas *3A perspektyva: anotavimas*, *abstrakcija*, *analizė* (Wallis, Nelson 2001: 311). Anotavimas yra dar vieno tekstynų skaidymo į rūšis kriterijus: tekstynai gali būti anotuoti ir neanotuoti. Kad tekstynai geriau tiktų moksliniams tyrimams, jie dažnai anotuojami: tekstas sužymimas nurodant jo struktūros elementus, vardo raidžių ar kitas santrumpas, akronimus, knygų įžangas, skyrius, pavadinimus, tikrinius vardus ir kt. Anotuojant sakytinę kalbą galima žymėti tarmines formas, kalbėjimą vienu metu, pauzes ir t. t. Žymint gramatinius duomenis daug įtakos turi subjektyvi tyrėjo nuomonė, nes dažniausiai naudojamos savitos, susitarimu paremtos gramatikos, skirtos kalbai apdoroti kompiuteriu (angl. *consensus grammar*). Todėl yra išsakytas ir neigiamas požiūris dėl anotavimo: gavęs anotuotą tekstą kompiuteris netikrina gramatinių kategorijų, bet priima jas tokias, kokios yra nurodytos, vadinasi, kompiuteris dirba su žymomis, o ne su pačia kalba (Marcinkevičienė 2000: 13). Gramatinis anotavimas išsamiau aptariamas tolesniuose poskyriuose.

Neanotuoti tekstynai naudojami kaip autentiškų duomenų šaltiniai. Rengiant naujus žodynus antraštinių žodžių sąrašas sudaromas remiantis tekstynuose pavartotų žodžių dažnumu. Pastaruoju metu naudojantis tekstyno duomenimis atnaujinami bei modernizuojami ir seniau parengti tradiciniai žodynai (Dadurkevičius, Petrauskaitė 2022: 123).

2.2. Morfologinis tekstynų anotavimas

Pradžioje tekstynai buvo sudaryti tik iš pačių žodžių, vėliau kiekvienam žodžiui buvo nurodyta, kokiai kalbos daliai jis priklauso. Anglų kalbai tai labai aktualu, nes joje dauguma žodžių gali turėti tiek daiktavardžio, tiek veiksmažodžio kategoriją. Todėl informacija apie kalbos dalį svarbi net ir atliekant naujausius tyrimus. Pavyzdžiui, kuriant sentimentų analizatorių, sakinyje *I like you* žodis *like* yra veiksmažodis ir turi pozityvų sentimentą, tačiau sakinyje *I am like you* šis žodis jau yra prielinksnis su neutraliu sentimentu (Ramachandran 2018). Iš esmės morfologinis anotavimas yra specialių žymų suteikimas kiekvienam žodžiui tekstyne, nurodant ne

[38] „Tekstynų lingvistikos pradžia laikomi 1961 m., kai Nelsonas Francis ir Henry Kučera anonsavo ir ėmėsi kurti garsųjį elektroninį rašytinės kalbos Browno tekstyną“ (Grumadienė 2002: 24).

tik tai, kokiai kalbos daliai jis priklauso, bet ir kitas morfologines kategorijas, tokias kaip laikas, skaičius, linksnis, asmuo, laipsnis ir pan. (23 interneto nuoroda[39]).

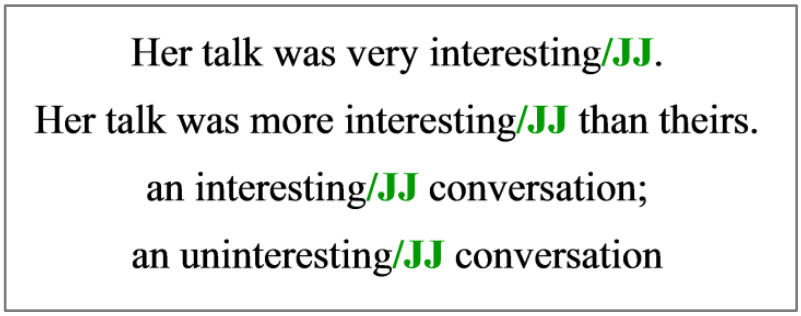
2.2.1. Pirmieji morfologiškai anotuoti tekstynai

Pirmajame anotuotame *Brauno tekstyne* (angl. *Brown Corpus*) visos pagalbinių veiksmažodžių formos turi atskiras žymas, pvz.: žodžio *be* žyma yra BE, o žodžio *am* žyma – BEM, žodžiui *been* nurodoma BEN ir kt. Analogiškai koduojamas ir žodis *do*: jis žymimas DO, forma *did* žymima DOD, forma *does* – DOZ ir pan. Žodis *have* taip pat turi atskiras žymas savo formoms: pagrindinei formai *have* naudojama žyma HV, vienaskaitos trečiojo asmens forma *has* žymima HVZ, žodis *having* – HVG ir t. t. Savarankiškos reikšmės veiksmažodžiai koduojami labiau apibendrintai: pagrindinė forma žymima VB (verb, base form), vienaskaitos trečiasis asmuo – VBZ (verb, 3rd present singular). Panašiai žymimi ir daiktavardžiai, pvz., bendrinių daiktavardžių vienaskaita žymima NN, daugiskaita – NNS (plural noun), tikriniams daiktavardžiams naudojamas kodas NP (proper noun), jų daugiskaitai – NPS (plural proper noun) ir pan. Tokiu pat būdu skiriama ir kitų kalbos dalių vienaskaita bei daugiskaita, pvz., apibrėžiamieji vienaskaitos įvardžiai turi žymą DT (singular determiner, pvz.: *this*, *that*), jų daugiskaita žymima DTS (plural determiner, pvz.: *these*, *those*). Klausiamiesiems ir santykiniams įvardžiams apibūdinti naudojamos dvi žymos: vardininko ir netiesioginio linksnio, t. y. WPS (nominative wh- pronoun, pvz.: *who*, *which*) ar WPO (objective wh- pronoun, pvz.: *whom, which*). Kitoms kalbos dalims dažniausiai nurodoma tik rūšis, pvz., kelintiniai skaitvardžiai žymimi OD (ordinal numeral, pvz.: *first*, *2nd*) ir t. t. (24 interneto nuoroda[40]).

Kuriant Pensilvanijos universitete (University of Pennsylvania) sintaksiškai anotuotą tekstyną *Penn Treebank* buvo atsisakyta žymų pertekliškumo ir visų veiksmažodžių formoms naudotas tas pats žymų rinkinys (Taylor, Marcus, Santorini 2003: 6). Žymos pateikiamos po žodžio, atskiriant jas pasviruoju brūkšniu, pvz., būdvardis žymimas JJ. Jo žymėjimo pavyzdžiai įvairiuose kontekstuose parodyti 19 pav. (Santorini 1991: 14).

[39] Prieiga internete: https://www.sketchengine.eu/pos-tags/ [žiūrėta 2022-11-22].

[40] Prieiga internete: https://en.wikipedia.org/wiki/Brown_Corpus#Part-of-speech_tags_used [žiūrėta 2022-11-22].



19 pav. Būdvardžio žymėjimo pavyzdžiai anotuojant tekstyną *Penn Treebank* (parengta pagal Santorini 1991: 14)

Tačiau net ir praėjus porai dešimtmečių, 2012 m. Rainhardas Kioleris (Reinhard Köhler) išsakė mintį, kad „nėra bendro vieningo standarto, kaip turi būti anotuojami tekstynai“[41] (Köhler 2012: 32). Žymų įvairovė ypač gerai matyti anotuojant iš principo skirtingas kalbas.

2015 m. buvo aprašytas pirmas morfologiškai anotuotas *Egipto arabų vaikų šnekamosios kalbos tekstynas*. Jį sudaro 10-ies vaikų nuo 1,7 iki 3,8 metų amžiaus įrašai, kurių apimtis 25 645 žodžiai (Salama 2020: 1). Autorė nurodo, kad arabų kalba labai skiriasi nuo kitų, tokių kaip anglų, prancūzų, vokiečių, japonų, kurioms jau sukurti morfologiniai anotatoriai. Todėl nebuvo galima pasinaudoti kitų šalių patirtimi ir reikėjo sukurti savas žymas. Arabų kalbininkai savo darbuose daugiausia dėmesio skyrė tradicinei arabų kalbos gramatikai, visai neatsižvelgdami į indoeuropiečių kalbų struktūrą. Arabų žodžius jie skirsto į tris pagrindines kalbos dalis: vardažodžius, veiksmažodžius ir dalelytes. Vardažodžiais vadinami žodžiai, kurie nusako asmenis, daiktus ar sąvokas. Veiksmažodžių klasifikacija panaši į anglų. Dalelytėms priklauso prielinksniai, prieveiksmiai, jungtukai, dalelytės, kiekiniai žodžiai, pertarai (angl. *filler*). Pertaras arabų kalboje – tai garsas ar žodis, kurį pokalbio metu ištaria vienas jo dalyvis, įspėdamas kitus, kad jis tik daro pauzę, norėdamas pamąstyti, bet dar nebaigė kalbėti. Tokie pavyzdžiai galėtų būti žodžių junginys *tai reiškia* arba keiksmažodžiai, kurie yra ypač dažnai vartojami kaip pertarai.

Tradiciškai arabų vardažodžiai skirstomi į darinius (žodžius, kurie padaryti iš veiksmažodžių, kitų vardažodžių, dalelyčių) ir pirminius, kurie nėra dariniai. Toliau jie kategorizuojami pagal skaičių, giminę ir linksnį. Ši klasė apima ir tuos žodžius, kurie indoeuropiečių kalbose laikomi dalyviais ir santykiniais, parodomaisiais bei

[41] “there is no general standard as to how corpora should be structured and notated” (Köhler 2012: 32).

klausiamaisiais įvardžiais. Veiksmažodžiai gali turėti įvykio veikslo esamąjį laiką (perfektą), eigos veikslo būtąjį laiką (imperfektą) ir liepiamąją nuosaką (imperatyvą). Toliau jie skirstomi pagal skaičių, asmenį ir giminę.

Pagrindinė kalbos dalies schema, naudojama anotuojant tekstyną (Salama, Alansary 2015: 2), parodyta 20 pav.

category: subcategory: subcategory

20 pav. Kalbos dalies schema, naudojama anotuojant *Egipto arabų vaikų šnekamosios kalbos tekstyną* (parengta pagal Salama, Alansary 2015: 2)

Anotuojant tekstyną kiekvienam žodžiui buvo nurodyta lema arba kamienas, ir tai jau traktuojama kaip morfeminės analizės dalis (Salama, Alansary 2015: 3–6).

Naujausiose morfologinės analizės sistemose informacija pateikiama labai populiariai: kalbos dalys parašomos be sutrumpinimų, kiekviena jų žymima vis kita spalva. 21 pav. parodyta angliško sakinio *John likes the blue house at the end of the street* analizė (6 interneto nuoroda[42], skirtukas *POS tagging*), atlikta su Stenfordo universiteto anotatoriumi *Stanford University Parts-Of-Speech Tagger* (5 interneto nuoroda, skirtukas *about Parts-of-speech.Info*). Kaip matyti iš paveikslėlio (21 pav.), autoriai teigia, kad „kompiuteriai taip pat padaro klaidų“, tačiau taip pasakyti galima tik apie kompiuterinį tautų kalbų apdorojimą. Atlikdami aritmetinius veiksmus ar kitas formaliai aprašytas užduotis, kompiuteriai klaidų nepadaro niekada.

21 pav. Anglų kalbos sakinio morfologinė analizė (6 interneto nuoroda)

[42] Prieiga internete: https://parts-of-speech.info/ [žiūrėta 2022-11-22].

2.2.2. Lietuvių kalbos morfologiškai anotuotas tekstynas MATAS

Pirmasis morfologiškai anotuotas lietuvių kalbos tekstynas MATAS (25 interneto nuoroda[43]) buvo kuriamas 2002–2014 m. Pradiniame etape jo pagrindą sudarė 1 mln. žodžių. Iš pradžių tekstai buvo anotuojami automatiškai, naudojant V. Zinkevičiaus sukurtą morfologinės analizės programinę įrangą (Zinkevičius 2000). Gautus rezultatus peržiūrėjo kalbininkai: sutvarkė neteisingai anotuotų žodžių žymas, prie morfologinio anotatoriaus neatpažintų žodžių įrašė trūkstamą informaciją. Vėliau tekstynas buvo papildytas iki 1,6 mln. žodžių. Tekstai imti iš keturių sričių: grožinės literatūros, periodikos, mokslinių ir administracinių leidinių (Bielinskienė, Boizou, Rimkutė 2017: 2–3). Internete jis laisvai prieinamas dviem formatais: CoNLL-U (Computational Natural Language Learning – Universal Dependencies) ir TAB-WPL (Tab delimited Word Per Line). Antrasis reiškia, kad į kiekvieną eilutę įrašomi duomenys apie vieną žodį ir duomenų tipai atskiriami tabuliacijomis. Pats tekstas užrašomas XML formatu, pvz., sakinio pradžia žymima <s>, o sakinio pabaiga – </s>. 2 priede parodytas šiuo formatu anotuoto teksto fragmentas. Pirmajame stulpelyje pateikiami sakinio žodžiai, antrajame – lema (pradinė žodžio forma: daiktavardžių, būdvardžių ir kt. vardininkas; veiksmažodžių bendratis ir t. t.), trečiajame stulpelyje – morfologiniai duomenys, užrašyti MULTEXT-East žymomis. Naudojant šį metodą anotavimo informacija surašoma į raidžių seką, kurioje kiekviena morfologinė kategorija nurodoma viena raide. Svarbu atkreipti dėmesį į tai, kad jos išdėstomos labai griežta tvarka: kiekviena kategorija turi savo poziciją, o jei kuriai nors kalbos daliai tam tikra kategorija yra nebūdinga (pvz., būdvardžių bevardei giminei – linksnis), jos vietoje rašomas brūkšnelis. Kategorijos reikšmė žymima anglų kalbos žodžių sutrumpinimais, pvz.: daiktavardis žymimas *n* (noun), veiksmažodis – *v* (verb), moteriškoji giminė – *f* (feminine) ir t. t. Visiems tos pačios kalbos dalies žodžiams skiriamas fiksuotas pozicijų skaičius: daiktavardžiams ir būdvardžiams – 7, veiksmažodžiams – 14, prieveiksmiams – 3 ir pan.

[43] Prieiga internete: https://klc.vdu.lt/matas-morfologiskai-anotuotas-tekstynas/ [žiūrėta 2022-11-22].

Pateikiant informaciją kitu formatu – CoNLL-U – naudojamos trijų tipų eilutės:

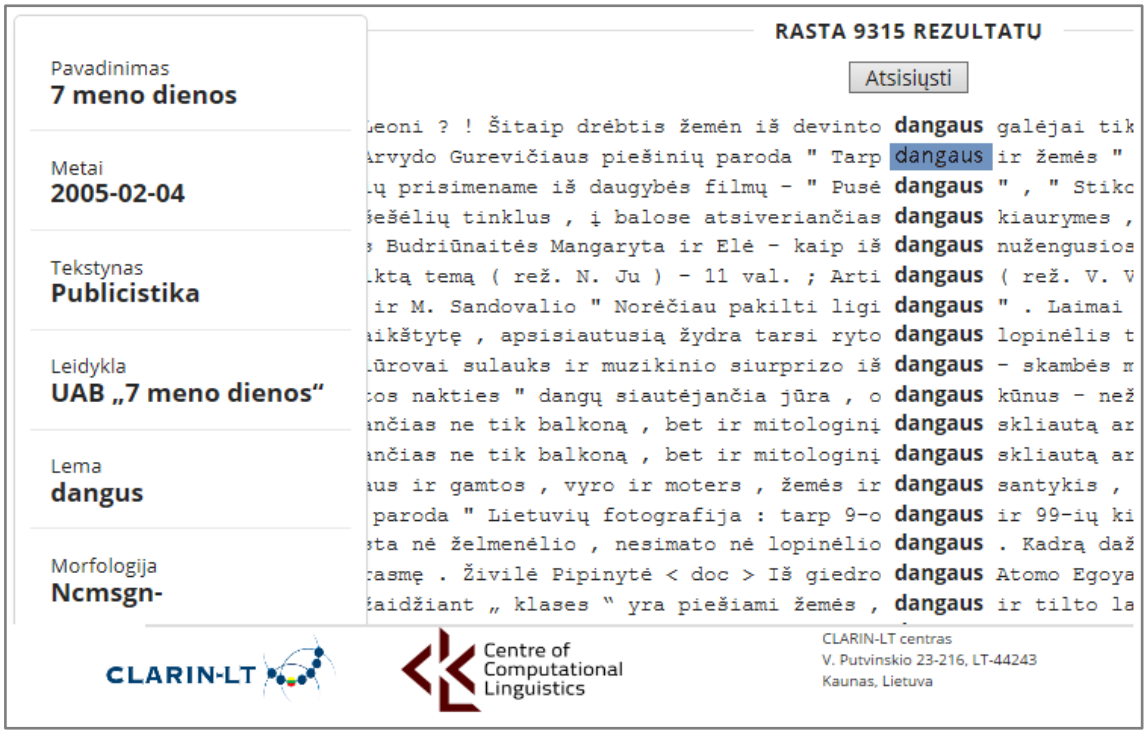
a) žodžio eilutė, kurioje nurodomi anotavimo duomenys, įrašyti į 10 stulpelių, atskirtų tabuliacijomis;
b) tuščia eilutė, žyminti sakinių ribas;
c) komentaro eilutė, prasidedanti grotelėmis (#).

Sakinys vaizduojamas kaip viena ar daugiau žodžio eilučių, kurios turi tokius laukelius:

1) žodžio eilės numeris, kuris kiekvienam sakiniui prasideda nuo 1;
2) sakinyje pavartota žodžio forma;
3) lema – pradinė žodžio forma;
4) kalbos dalis, kuriai priklauso žodis;
5) specifinis, lietuvių kalbai būdingų morfologinių žymų kodas – JABLONSKIS (plačiau žr. Terminų žodynėlyje);
6) morfologinės žymos *Universal dependencies* pagrindu (26 interneto nuoroda[44]); kiekviena žyma turi formą *Name=Value*; žymos atskiriamos viena nuo kitos vertikaliu brūkšniu, pvz., *Gender=Masc | Number=Sing*;
7) 7-asis, 8-asis ir 9-asis stulpeliai tekstyne MATAS nenaudojami;
8) 10-ajame stulpelyje po raktažodžio MULTEXT nurodomos morfologinės žymos šiuo formatu.

3 priede pateikiamas tas pats teksto fragmentas, kaip ir 2 priede, tik anotuotas CoNLL-U formatu. MULTEXT-East žymas galima matyti ir atliekant paiešką tekstyne (27 interneto nuoroda[45]), pažymėjus ieškomą žodį. 22 pav. pateikiamas žodžio *dangaus*, esančio antroje eilutėje, pavyzdys. Kadangi MULTEXT-East žymos labiau skirtos kompiuteriniam kalbos apdorojimui, todėl buvo sukurtos lietuvių kalbai specifinės žymos JABLONSKIS, su kuriomis patogiau dirbti lietuvių kalbininkams. Naudojant lietuviškų morfologinių kategorijų sutrumpinimus tikimasi minimalizuoti žmogaus daromas klaidas.

[44] Prieiga internete: https://universaldependencies.org/u/overview/morphology.html [žiūrėta 2022-11-22].
[45] Prieiga internete: http://corpus.vdu.lt [žiūrėta 2022-11-22].



22 pav. Žodžio *dangaus* paieškos tekstyne rezultatų pavyzdys (28 interneto nuoroda[46])

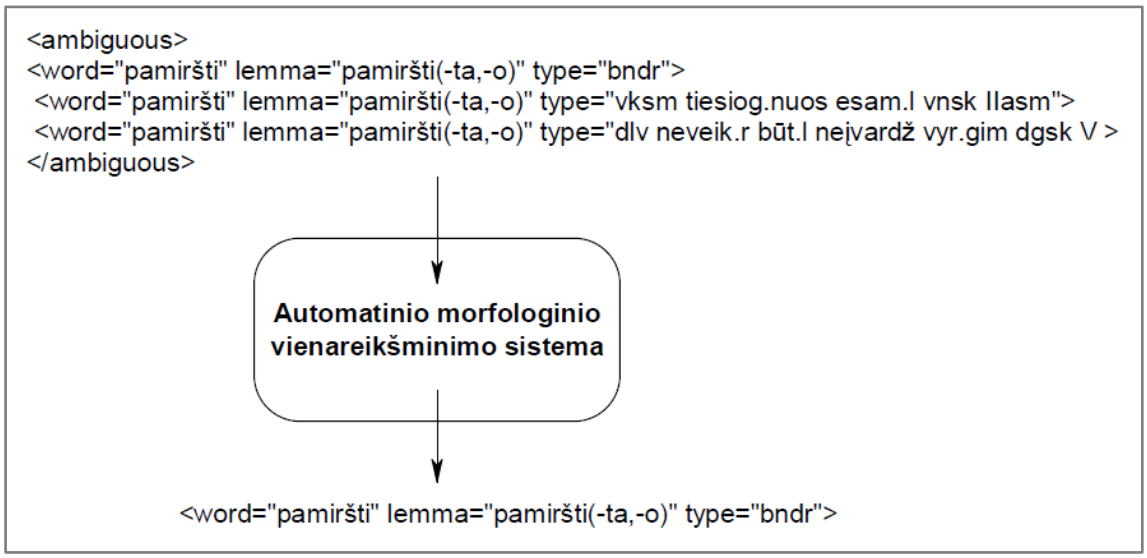
2.2.3. Statistinių metodų pagrindu anotuotas lietuvių kalbos tekstynas

2006 m. VDU buvo morfologiškai anotuotas 100 mln. žodžių apimties tekstynas, naudojant statistinius metodus, kurių dėka turėjo būti panaikintas morfologinis daugiareikšmiškumas. Priešingai nei MATAS, šis tekstynas nėra viešai prieinamas. Tačiau verta aptarti metodiką, kaip, siekiant patobulinti taisyklių pagrindu veikiantį lietuvių kalbos morfologinį analizatorių, buvo pasitelkti statistiniai metodai. V. Zinkevičiaus sukurta lietuvių kalbos morfologinės analizės ir sintezės programinė įranga pateikia visus galimus žodžio morfologinius variantus, pvz., įvedus raidžių seką *mes*, nurodoma, kad tai įvardis, daugiskaitos vardininkas arba veiksmažodžio *mesti* būsimojo laiko trečiasis asmuo. 1 mln. žodžių apimties tekstynas buvo vienareikšminamas ranka ir tam prireikė penkerių metų vieno žmogaus darbo. Norint paspartinti šį procesą, nuspręsta jį automatizuoti pasitelkus statistinius metodus. Buvo pasinaudota čekų patirtimi, kai statistiniai metodai jungiami su taisyklėmis pagrįstu metodu. Tokia metodika yra nepriklausoma nuo kalbos, todėl galėjo būti taikoma ir lietuvių kalbos tekstams. Vienintelis nuo kalbos priklausantis dalykas yra nedidelis

[46] Prieiga internete: http://corpus.vdu.lt/lt/?word=dangaus [žiūrėta 2022-11-22].

ranka morfologiškai anotuotas tekstynas – tai mokymo duomenys, skirti statistinių metodų pagrindu veikiančiai analizatoriaus daliai (Rimkutė, Daudaravičius 2007: 31).

Iš pradžių atliekama morfologinė analizė, naudojant programinę įrangą, apdorojančią tekstus taisyklėmis pagrįstu metodu. Vėliau tie žodžiai, kuriems analizatorius nurodo po kelis galimus morfologinių duomenų variantus, vienareikšminami. Automatinio morfologinio vienareikšminimo sistemai, veikiančiai tikimybinių metodų pagrindu, pateikiami visi gauti žodžio morfologinių požymių variantai ir ji išrenka iš jų vieną labiausiai tikėtiną. Morfologinio vienareikšminimo procesas (Rimkutė, Daudaravičius 2007: 32) parodytas 23 pav.



23 pav. Morfologinio vienareikšminimo procesas (Rimkutė, Daudaravičius 2007: 32)

Naudojant šią morfologinio anotavimo metodiką buvo pasiektas 94 proc. tikslumas. Vadinasi, ne visus daugiareikšmiškumo atvejus pavyksta išspręsti. Tačiau net ir kalbininkas be platesnio konteksto ne visada gali suprasti, kuri forma – veiksmažodis ar daiktavardis – pavartota, pvz., *kovos dėl teisės likti pirmajame ešelone*, *kovos su narkotikais*. Net ir nusprendus, kad *kovos* yra daiktavardis, lieka neaišku, ar tai vienaskaitos kilmininkas, ar daugiskaitos vardininkas. Čia galėtų pagelbėti tik platesnis kontekstas (Rimkutė, Daudaravičius 2007: 33), kuris automatinio kalbos apdorojimo metu būna gana ribotas. Kitų šalių mokslininkai išsako panašias mintis: „Iki šiol dar negalima pasiekti, kad automatinis anotavimas būtų išsamus ir be klaidų“[47] (Sasaki, Witt 2004: 199).

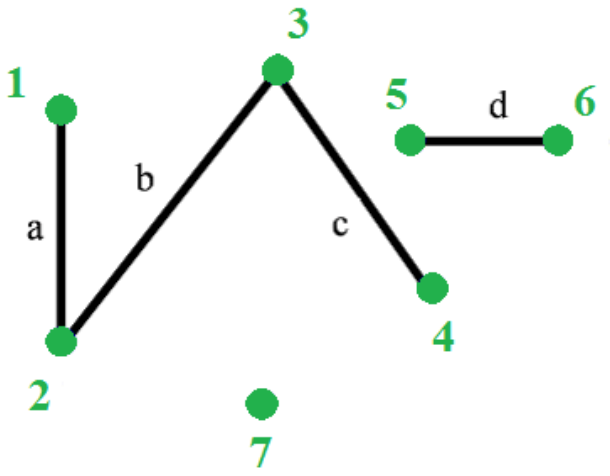
[47] „Es ist bisher nicht möglich, eine sowohl fehlerfreihe, als auch vollständige linguistische Annotation maschinell su erreichen“ (Sasaki, Witt 2004: 199).

2.3. Sintaksinis tekstynų anotavimas

Anotuoti tekstynai, kuriuose pateikiami ne tik žodžių morfologiniai požymiai, bet ir sakinių sintaksinės struktūros, angliškai vadinami *treebank*, nes „labiausiai paplitęs sakinio struktūros pavaizdavimo būdas yra medis“[48] (Allen 1987: 41), tiksliau, grafo medis. Tai nėra lingvistinės, kasdien kalbininkų vartojamos sąvokos, todėl čia pateikiama šiek tiek informacijos iš grafų teorijos.

2.3.1. Medis grafų teorijoje

Pagrindinis medžio apibrėžimas skamba taip: „Medis – tai susietas grafas be ciklų“[49] (Swamy, Thulasiraman 1981: 32). Šiame apibrėžime minimi keturi terminai: *grafas*, *ciklas*, *medis* ir *susietas*. Iš pradžių reikėtų išsiaiškinti, kas vadinama *grafu*. Grafas – tai grupė objektų, sujungtų linijomis, kurios vadinamos lankais. Patys objektai vadinami viršūnėmis. Viršūnės dažniausiai numeruojamos, o lankams suteikiami raidžių pavadinimai. Grafas gali būti *susietas* arba *nesusietas*. Nesusieto grafo pavyzdys pateikiamas 24 pav. „Grafas vadinamas susietu, jei tarp bet kurių dviejų jo viršūnių egzistuoja kelias“ (Swamy, Thulasiraman 1981: 13), t. y. jei iš kiekvienos viršūnės einant linijomis galima patekti į visas kitas grafo viršūnes.



24 pav. Nesusieto grafo pavyzdys

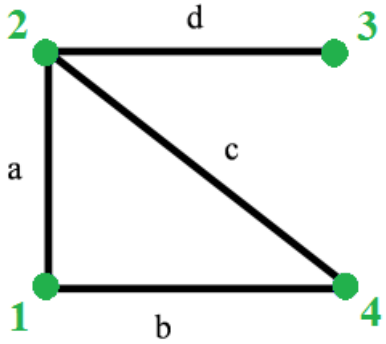
[48] “The most common way of representing a sentence structure is to use a treelike representation” (Allen 1987: 41).

[49] “A tree is a connected acyclic graph” (Swamy, Thulasiraman 1981: 32).

24 pav. pavaizduotas grafas turi viršūnę *7*, iš kurios neišeina nė viena linija, vadinasi, iš šios viršūnės negalima patekti į jokią kitą viršūnę. Iš viršūnės *5* galima patekti į viršūnę *6*, einant linija *d*, bet iš jos (viršūnės *5*) negalima patekti į viršūnę *7* ar *3*, taigi, šis grafas netenkina susieto grafo apibrėžimo sąlygų, todėl yra nesusietas.

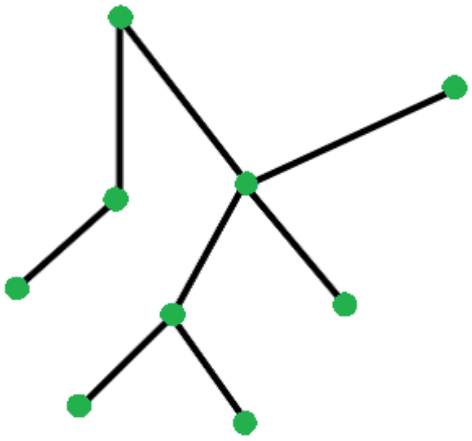
Susietas grafas parodytas 25 pav.: čia iš kiekvienos viršūnės yra kelias į bet kurią kitą grafo viršūnę, t. y., jei mes esame kurioje nors viršūnėje, eidami linijomis, galime patekti į visas kitas šio grafo viršūnes.

Dar vienas neįprastas lingvistikoje terminas, vartojamas medžio apibrėžime, yra *ciklas*. Sakoma, kad grafas turi ciklą (ar kelis ciklus), jei, išėjus iš kurios nors viršūnės, į ją galima sugrįžti kitu keliu. Išėjus iš viršūnės *4* (žr. 25 pav.), einant linija *c*, galima patekti į viršūnę *2*, o sugrįžti jau kitu keliu – linijomis *a* ir *b*, vadinasi, 25 pav. pavaizduotas grafas turi ciklą.



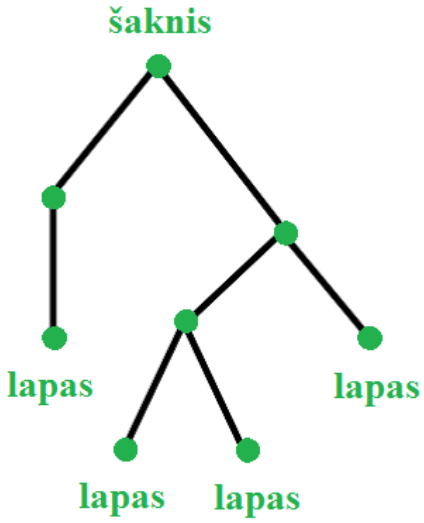
25 pav. Susieto grafo pavyzdys

Pagal apibrėžimą medis negali turėti ciklų. Medžio pavyzdys pateiktas 26 pav. Iš bet kurios viršūnės išėjus į kurią nors kitą, sugrįžti galima tik tuo pačiu keliu, kuriuo buvo eita, kito kelio atgal nėra.



26 pav. Grafų teorijos medžio pavyzdys

Lingvistikoje medžiai visada braižomi viršūne žemyn (žr. 27 pav.), t. y. jų šaknis yra viršuje, o lapai – apačioje (Batori, Lenders, Putschke 1989: 23) ir linijos eina visada iš viršaus į apačią. Čia reikėtų aptarti dar du grafų teorijos terminus *šaknis* ir *lapas*. Šaknis yra tokia grafo viršūnė, į kurią neateina jokia linija, iš jos linijos gali tik išeiti. Lapas – tai tokia viršūnė, į kurią linija ateina, bet iš jos toliau nebeišeina nė viena linija.



27 pav. Kalbininkų medžio pavyzdys

Tačiau patys pirmieji sintaksiškai anotuoti tekstynai dar nenaudojo medžių sakinio struktūrai pavaizduoti.

2.3.2. Sintaksinio anotavimo formatai

2000 m. Anglijoje, Sasekso universitete (University of Sussex), *Brauno tekstyno* (29 interneto nuoroda[50]) pagrindu buvo sukurtas sintaksiškai anotuotas tekstynas SUSANNE (30 interneto nuoroda[51]). Jame informacija pateikiama lentelėje, duomenis išdėstant į šešis stulpelius. Pirmajame stulpelyje užrašomas unikalus eilutės numeris tekstyne. Antrajame – nurodomas žodžio statusas: raidė *A* reiškia santrumpą, raidė *S* – simbolį, raidė *E* – korektūros klaidą, tačiau daugumai žodžių rašomas brūkšnelis. Trečiajame stulpelyje pateikiama morfologinė informacija naudojant Lankasterio universiteto (Lancaster University) žymas[52] (31 interneto nuoroda[53]). Ketvirtajame

[50] Prieiga internete: https://en.wikipedia.org/wiki/Brown_Corpus [žiūrėta 2022-11-22].
[51] Prieiga internete: https://www.grsampson.net/SueDoc.html [žiūrėta 2022-11-22].
[52] Lancaster University Tagset UCREL CLAWS7.
[53] Prieiga internete: http://ucrel.lancs.ac.uk/claws7tags.html [žiūrėta 2022-11-22].

stulpelyje įrašomas pats sakinyje pavartotas žodis, penktajame – jo lema, šeštajame – sintaksiniai duomenys (28 pav.).

```
N06:0180.12   -   NN1u    Baldness   baldness  [S[Ns:s.Ns:s]
N06:0180.15   -   VBDZ    was        be        [Vsu.
N06:0180.18   -   VVGt    attacking  attack    .Vsu]
N06:0180.21   -   APPGm   his        he        [Ns:o.
N06:0180.24   -   NN1c    pate       pate      .Ns:o]S]
```

28 pav. SUSANNE tekstyno pavyzdys

Vokietijoje, Zarlando universitete, sukurtas sintaksiškai anotuotas tekstynas *NEGRA Corpus*. Tekstyno failą sudaro anotuoti sakiniai, atskirti vienas nuo kito sakinio pradžios #BOS (Beginning Of Sentence) ir sakinio pabaigos #EOS (End Of Sentence) žymomis (32 interneto nuoroda[54]). 29 pav. pateikiamas sakinio *Sie gehen gewagte Verbindungen und Risiken ein, versuchen ihre Möglichkeiten auszureizen* analizės pavyzdys, kuriame naudojamas mišrus kodavimas: iš pradžių išdėstomi sakinio žodžiai kiekvienam skiriant po eilutę ir į stulpelius įrašant morfologinius duomenis, o po jų, apačioje, pateikiami ryšiai tarp sakinio žodžių, t. y. sintaksinė informacija (Köhler 2012: 41).

```
#BOS 2 2 899973978 1
Sie              PPER    3.Pl.*.Nom      SB      504
gehen            VVFIN   3.Pl.Pres.Ind   HD      504
gewagte          ADJA    Pos.*.Akk.Pl.St NK      500
Verbindungen     NN      Fem.Akk.Pl.*    NK      500
und              KON     --              CD      502
Risiken          NN      Neut.Akk.Pl.*   CJ      502
ein              PTKVZ   --              SVP     504
,                $,      --              --      0
versuchen        VVFIN   3.Pl.Pres.Ind   HD      505
ihre             PPOSAT  *.Akk.Pl        NK      501
Möglichkeiten    NN      Fem.Akk.Pl.*    NK      501
auszureizen      VVIZU   --              HD      503
.                $.      --              --      0
#500             NP      --              CJ      502
#501             NP      --              OA      503
#502             CNP     --              OA      504
#503             VP      --              OC      505
#504             S       --              CJ      506
#505             S       --              CJ      506
#506             CS      --              --      0
#EOS 2
```

29 pav. Tekstyno *NEGRA Corpus* pavyzdys (Köhler 2012: 41)

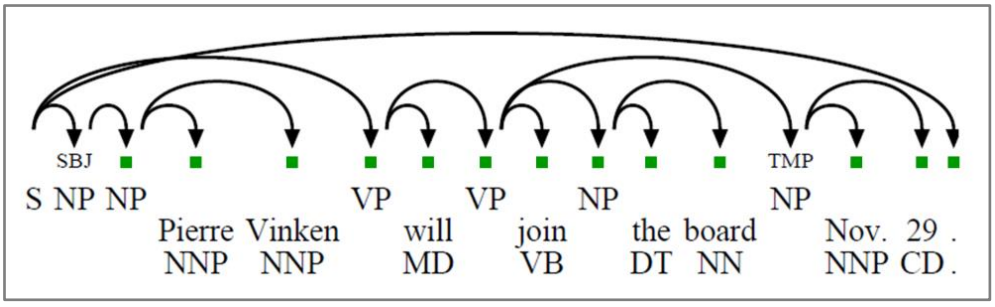
[54] Prieiga internete: https://github.com/nltk/nltk/issues/137 [žiūrėta 2022-11-22].

Vėlesni anotavimo formatai jau primena sintaksinės struktūros medį, nors dar koduojami tekstu (skliaustų hierarchija rodo žodžio rangą sakinyje ir pasukus popieriaus lapą 90° kampu galima įžiūrėti sakinio medį). Pensilvanijos universitete buvo parengta skliaustais koduota tekstyno versija *Penn Treebank*. 30 pav. parodytas sakinio *Jones followed him into the front room, closing the door behind him* pavyzdys, anotuotas *Penn Treebank* (Taylor, Marcus, Santorini 2003: 10).

```
( (S (NP-SBJ-1 Jones)
     (VP followed
         (NP him)
         (PP-DIR into
                 (NP the front room))
         ,
         (S-ADV (NP-SBJ *-1)
                (VP closing
                    (NP the door)
                    (PP behind
                        (NP him)))))
     .))
```

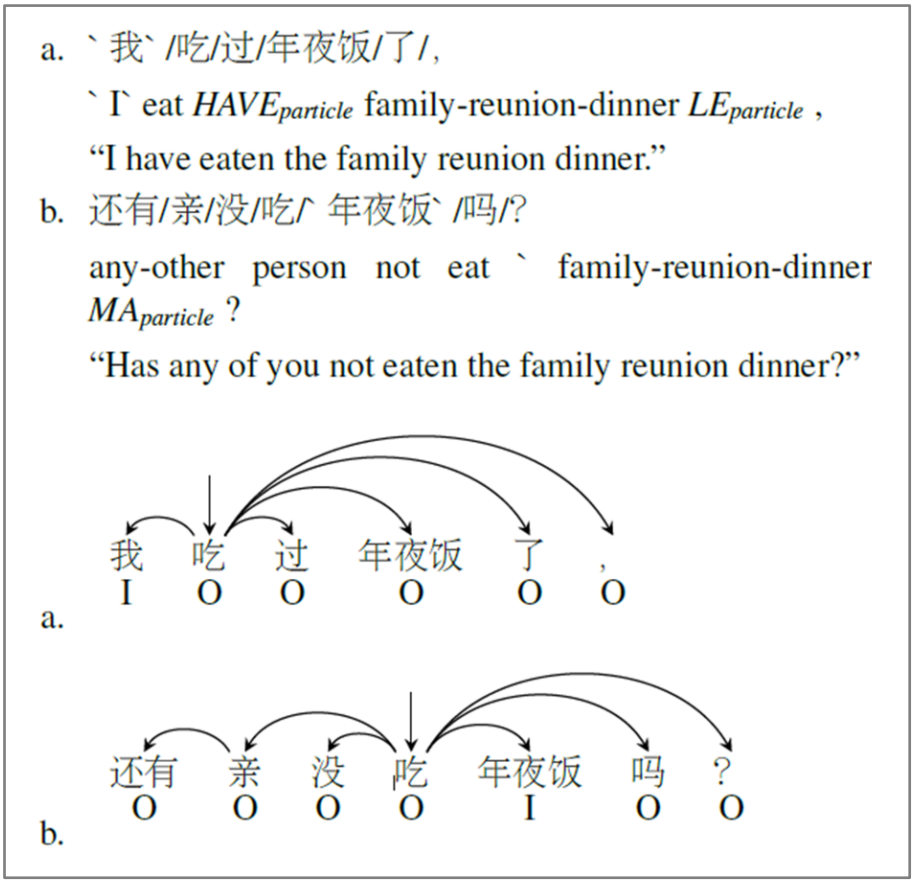
30 pav. Skliaustais koduoto sintaksinio anotavimo pavyzdys (Taylor, Marcus, Santorini 2003: 10)

Kopenhagos sintaksiškai anotuotame tekstyne (angl. *Copenhagen Dependency Treebank*) sintaksinė struktūra pateikiama daug vaizdžiau nei skliaustų metodu koduota. Jame ryšiai tarp žodžių nurodomi linijomis virš sakinio. Toks vaizdavimo būdas vadinamas lankų grafo formatu (angl. *arc-graph format*). Naudojant frazių gramatiką (apie ją plačiau žr. 4.1 poskyryje), sakinys pavaizduojamas grafu, išdėstant jo viršūnes vienoje linijoje ir papildomai dar įterpiant frazėms skirtas pozicijas atitinkamose vietose. 31 pav. pateiktas šiuo metodu anotuoto sakinio *Pierre Winken will join the board Nov. 29.* pavyzdys (Buch-Kromann 2010: 2).



31 pav. Lankų grafo metodu anotuoto sakinio pavyzdys iš *Kopenhagos sintaksiškai anotuoto tekstyno* (parengta pagal Buch-Kromann 2010: 2)

Panašiai anotuojami ir kinų kalbos sakiniai. 32 pav. pateikiamas sakinys iš kinų kalbos tekstyno, kuriame numatytas ir eliptinių sakinių apdorojimas. Ženklu „I“ žymimi atkurti žodžiai, kurie eliptiniame sakinyje buvo praleisti, o ženklas „O“ rodo tekste esančius žodžius (Ren ir kt. 2018: 1751).

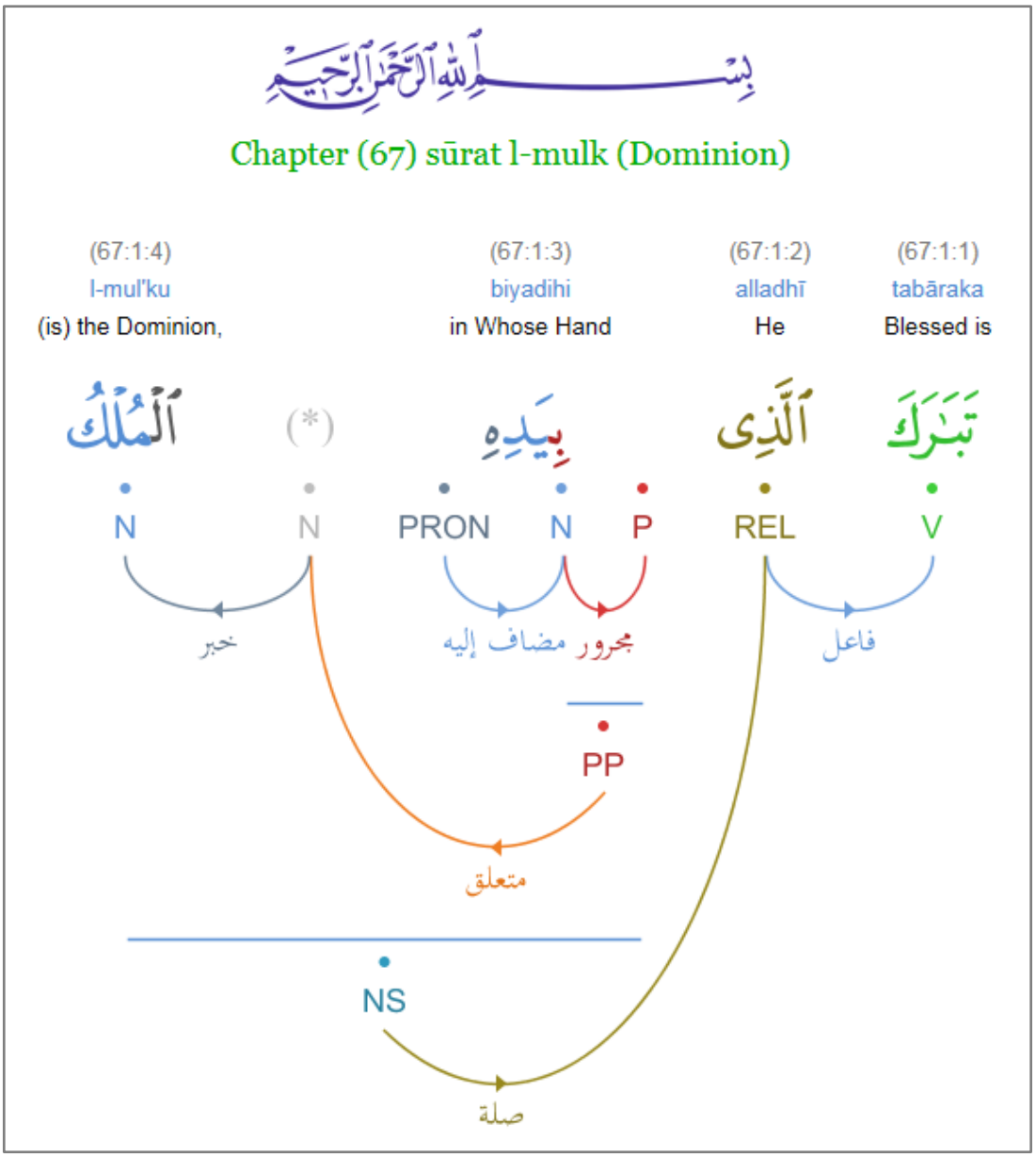


32 pav. Kinų kalbos tekstyno pavyzdys (Ren ir kt. 2018: 1751)

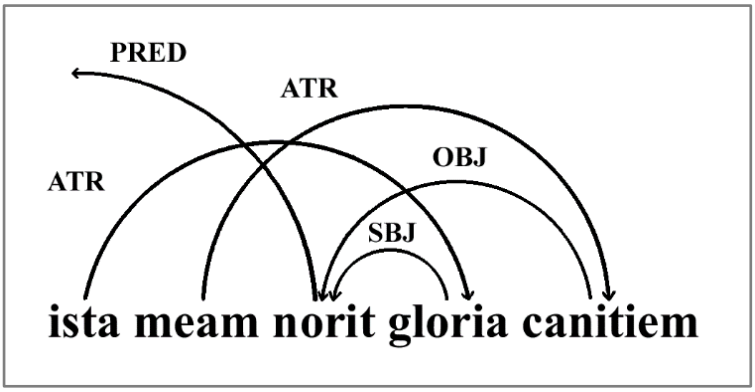
Arabų Korano tekstyne (angl. *Quranic Arabic Corpus*) ryšiams tarp žodžių parodyti taip pat naudojamos linijos, tik jos išdėstomos kiek kitaip – žodžio apačioje. Tarp žodžio ir linijos įterpiama informacija apie kalbos dalį. 33 pav. pateiktas šio tekstyno sakinys (33 interneto nuoroda[55]).

Tačiau grafo lankų formatu anotuoto lotynų kalbos sakinio struktūra atrodo labai paini. 34 pav. pateiktas *Senovės graikų ir lotynų kalbų tekstyno* pavyzdys (Bamman, Crane 2011: 81). Rodyklės išeina iš išplečiančio žodžio ir eina į išplečiamąjį. Virš jų nurodyti sintaksiniai ryšiai. Tačiau sakinio struktūroje matyti labai daug susikertančių linijų, todėl atsekti, kuris žodis nuo kurio priklauso, sudėtinga.

[55] Prieiga internete: http://corpus.quran.com/treebank.jsp?chapter=67 [žiūrėta 2022-11-22].

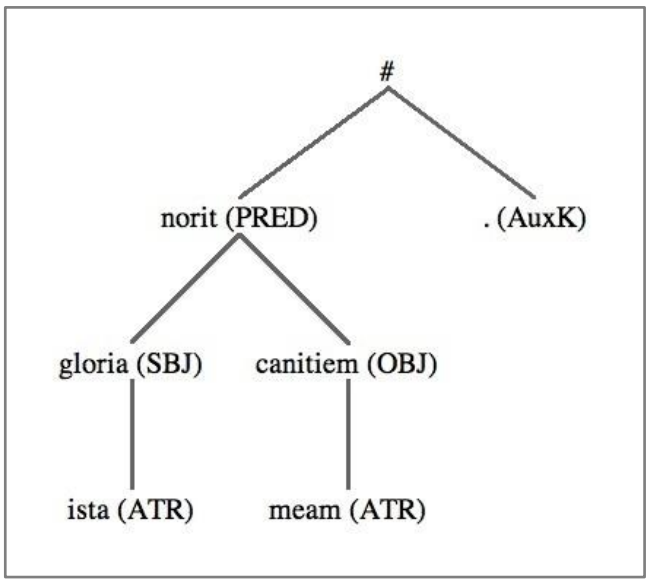


33 pav. *Arabų Korano tekstyno* pavyzdys (33 interneto nuoroda)



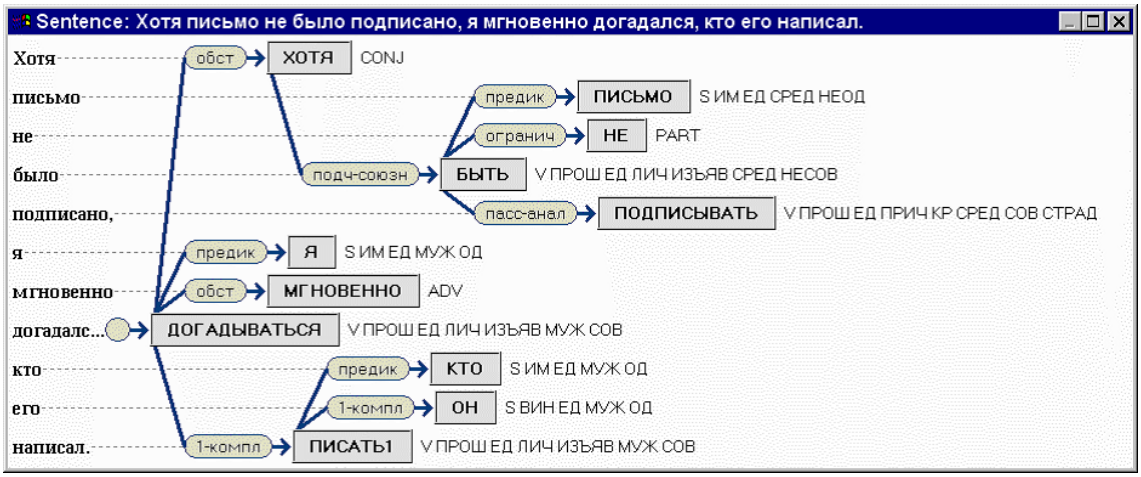
34 pav. Lotynų kalbos sakinio sintaksinė struktūra (Bamman, Crane 2011: 81)

Senovės graikų ir lotynų kalbų tekstyne buvo naudojamas dar ir kitas anotavimo formatas, kur jau galima matyti medžio vaizdą. 35 pav. pateikiamas to paties sakinio *Ista meam norit gloria canitiem* pavyzdys, kuriame šalia žodžio nurodoma jo sintaksinė funkcija (34 interneto nuoroda[56]).



35 pav. *Senovės graikų ir lotynų kalbų tekstyno* pavyzdys (34 interneto nuoroda)

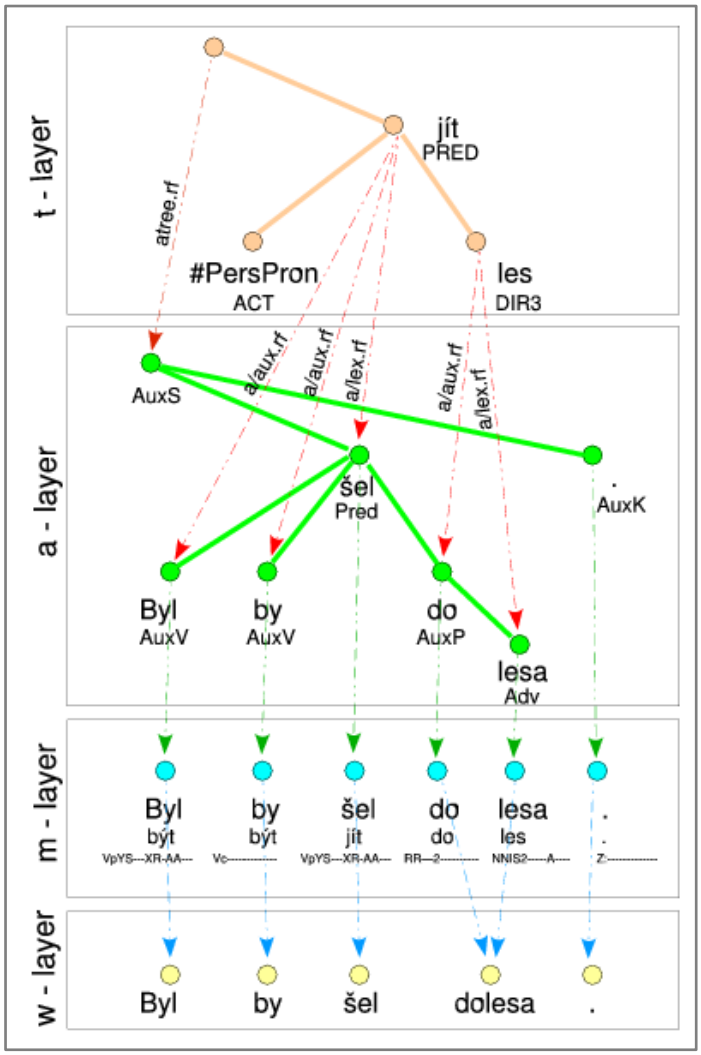
Rusų kalbos sintaksiškai anotuotame tekstyne sintaksinės kategorijos surašytos ant linijų, jungiančių priklausomus žodžius (Boguslavsky ir kt. 2000: 5). Sakinio *Хотя письмо не было подписано, я мгновенно догадался, кто его написал* analizė pavaizduota 36 pav.



36 pav. Rusų kalbos sintaksiškai anotuoto tekstyno pavyzdys (Boguslavsky ir kt. 2000: 5)

[56] Prieiga internete: http://nlp.perseus.tufts.edu/syntax/treebank/ [žiūrėta 2012-12-20].

Vėliau buvo pradėtas naudoti dar vienas formatas, įtraukiant ir sakinio semantikos duomenis. 37 pav. parodytas *Prahos sintaksiškai anotuoto tekstyno* (angl. *Prague Dependency Treebank*) sakinys.



37 pav. *Prahos sintaksiškai anotuoto tekstyno* pavyzdys (35 interneto nuoroda)

Informacija pateikiama keturiais lygmenimis (35 interneto nuoroda[57]):

a) žodžių lygmenyje (*w-layer*) įrašomi paties sakinio žodžiai,

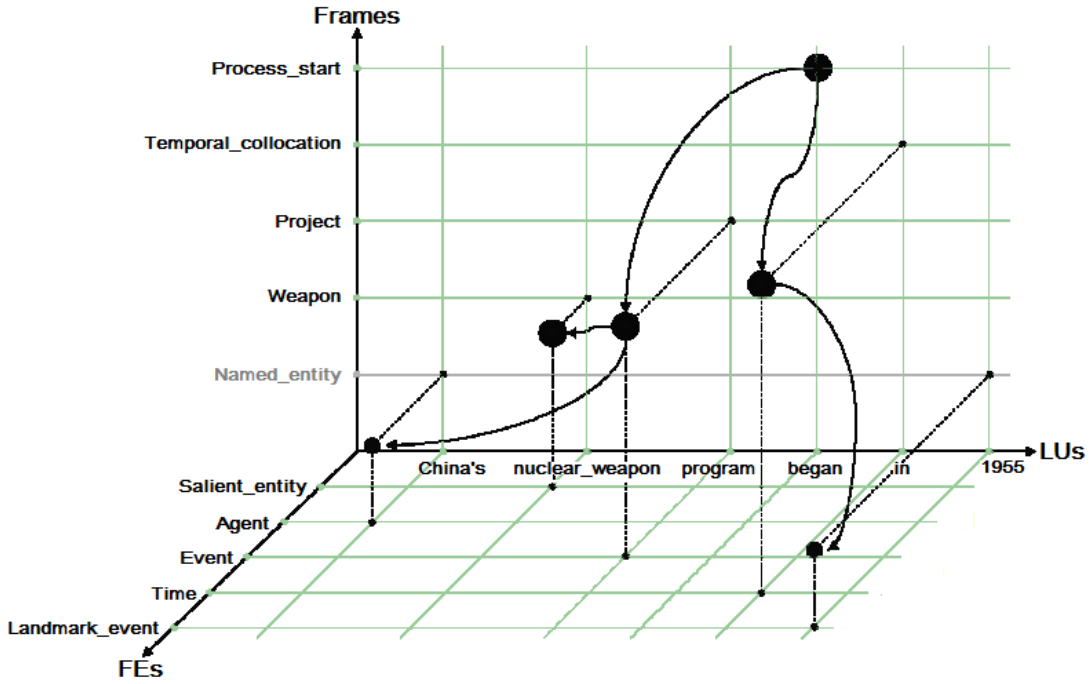
b) morfologiniame (*m-layer*) – jų pradinė forma (lema),

c) sintaksiniame (*a-layer*) – braižomas sakinio struktūros medis ir šalia žodžių nurodoma jų sintaksinė funkcija,

d) semantiniame (*t-layer*) – informacija taip pat pateikiama medžiu, tik čia nurodomos jau ne sintaksinės, bet semantinės funkcijos.

[57] Prieiga internete: http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch02.html [žiūrėta 2022-11-22].

Dar vienas būdas pavaizduoti sakinį, įtraukiant ir semantikos duomenis, – erdvinis (3D) formatas. Jis buvo naudotas latvių kalbininkų darbuose (Barzdiņš ir kt. 2008). Šiuo metodu anotuotas sakinys *China's nuclear weapon program began in 1955* parodytas 38 pav.



38 pav. Erdvinis (3D) anotavimo metodas (Barzdiņš ir kt. 2008: 280)

Siekiant suvienodinti sintaksinio anotavimo formatus, buvo sukurtos universaliosios priklausomybės (angl. *universal dependencies*). Apie tai plačiau žr. 4.1 poskyryje. Sintaksiškai anotuojant sakinius lietuvių kalbos tekstyne buvo taikomas Prahoje naudotas metodas PML – *Prague Mark-Up Language*.

2.3.3. Lietuvių kalbos sintaksiškai anotuotas tekstynas ALKSNIS

VDU Kompiuterinės lingvistikos centre buvo parengtas sintaksiškai anotuotas tekstynas ALKSNIS (anotuotas lietuvių kalbos sintaksinis tekstynas). 2016 m. internete pateiktoje pirmojoje versijoje anotuoti 2 355 sakiniai iš bendrosios bei specialiosios periodikos, grožinės literatūros ir administracinės srities. Rengiant tekstyną, imti ištisi, nesutrumpinti tekstai. Periodikos dalį sudaro straipsniai iš Lietuvoje leidžiamų laikraščių ir žurnalų, grožinės literatūros dalį – mažesni prozos

tekstai: trumpi apsakymai, esė. Visi anotuoti tekstai buvo publikuoti 2004–2014 m. (Bielinskienė ir kt. 2016: 108).

Anotuojant buvo pasitelktas lietuvių kalbos sintaksinis analizatorius, parengtas Čekijoje sukurtą programinę įrangą pritaikius lietuvių kalbai, todėl ir sintaksinės struktūros generuojamos remiantis čekų kalbai naudotu PML formatu. Kiekviena medžio viršūnė atitinka sakinio žodį, skyrybos ženklą ar kitą sakinio vienetą (simbolį, skaitmenį ir pan.). Prie kiekvieno žodžio nurodoma tokia informacija:

a) pats sakinio žodis,
b) jo pradinė forma (lema),
c) morfologiniai duomenys (kalbos dalis, giminė, skaičius, linksnis, asmuo, laikas ir kt.),
d) sintaksinė funkcija (sakinio dalis, t. y. tarinys, pažyminys ir t. t.).

Sintaksiniai ryšiai tarp žodžių parodomi lankais.

Čekų kalbos anotavimo metodika buvo pasirinkta remiantis lietuvių ir čekų kalbų panašumu. Sakiniui aprašyti naudojamos penkios sintaksinės funkcijos (Bielinskienė ir kt. 2016: 112):

1) predikatas,
2) subjektas,
3) objektas,
4) atributas,
5) modifikatorius.

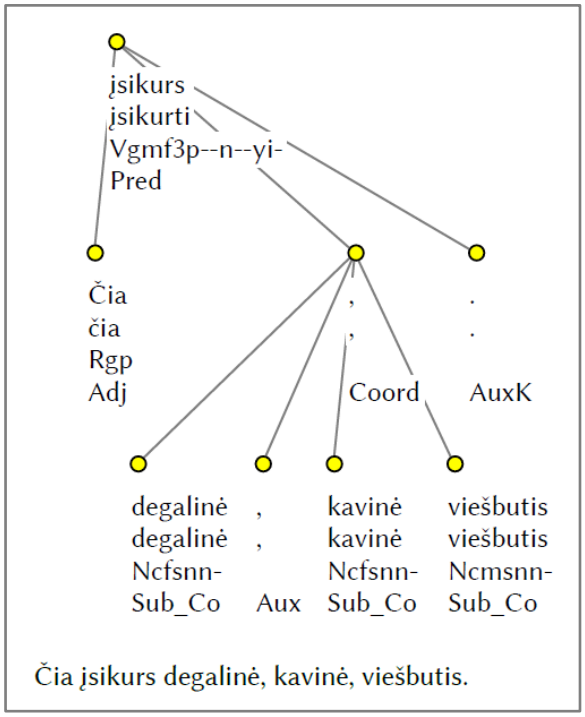
Jos žymimos atitinkamai: *Pred* (*PredN*, *PredV*), *Sub*, *Obj, Atr* ir *Adj*.

Šiuo metodu anotuotas sakinys *Čia įsikurs degalinė, kavinė, viešbutis* pavaizduotas 39 pav. (36 interneto nuoroda[58]). Visus automatiškai anotuotus sakinius peržiūri kalbininkų grupė ir ranka ištaiso pastebėtas analizatoriaus klaidas (37 interneto nuoroda[59]).

2019 m. buvo parengta ALKSNIS 3.0 versija ir tekstynas papildytas iki 3 563 sakinių apimties. Šioje versijoje informacija apie vieną sakinį pateikiama dviejuose failuose: PML ir CoNLL-U. Pastarasis apima visus morfologinius duomenis ir sintaksines funkcijas, todėl PML faile nubraižyta sintaksinė struktūra yra medis, kuriame prie kiekvienos viršūnės nurodyta tik lema. 4 priede pavaizduotas sakinys

[58] Prieiga internete: https://klc.vdu.lt/alksnis-sintaksiskai-anotuotas-tekstynas/ [žiūrėta 2022-11-22].
[59] Prieiga internete: http://clarin-lt.lt/?p=205 [žiūrėta 2022-11-22].

Kauno rajone prie magistralės „Via Baltica“ vakar atidarytas pirmasis pretenduojančio tapti didžiausiu Baltijos šalyse logistikos komplekso statinys, anotuotas ALKSNIS 3.0 versijos PML formatu, o 5 priede pateiktas tas pats sakinys CoNLL-U formatu.



39 pav. Sakinys iš Alksnis 2.1 versijos tb1-2-V1.pdf, 4 p. (36 interneto nuoroda)

2.4. Skyriaus išvados

Kai kurie tekstynai buvo kuriami kaip priemonė tam tikriems darbams atlikti ir viešai yra neprieinami (pirmasis lietuvių kalbos tekstynas, statistiniais metodais morfologiškai anotuotas *Dabartinės lietuvių kalbos tekstynas* ir kt.).

Pradėjus anotuoti tekstynus paaiškėjo, kad žymoms didelį poveikį turi kalbų įvairovė, todėl iki šiol nėra sukurto vienodo anotavimo standarto. Morfologiškai anotuotame lietuvių kalbos tekstyne šalia tarptautiniu mastu vartojamų angliškų žymų pateikiamos ir specialiai lietuvių kalbai sukurtos žymos iš lietuviškų morfologinių kategorijų santrumpų.

Sintaksiniam anotavimui atlikti buvo sukurta daug įvairių formatų, tačiau ne visi jie vienodai paplito. Bene populiariausias – Prahos mokslininkų pasiūlytas metodas. Daugelyje kalbų jis buvo pritaikytas savo reikmėms, taip pat naudotas ir anotuojant lietuvių kalbos sakinius. Mažiausiai pasiteisino erdvinis sakinio vaizdavimas – jis naudojamas nedaug kur.

3. MORFOLOGIJOS KOMPIUTERIZAVIMAS

Morfologinė analizė yra ypač svarbi fleksinėms kalboms, nes tai – pats pirmasis ir būtinas etapas atliekant bet kokią teksto analizę (Paikens 2007: 235). Morfologinis analizatorius gali suteikti daug vertingos informacijos atliekant ir kitus kompiuterinės lingvistikos darbus: lemavimą, sintaksinę analizę, automatinį vertimą, informacijos išgavimą iš teksto, tekstų grupavimą ir kt. (Tang 2006: 35).

3.1. Morfologiniai analizatoriai

Visus morfologinius analizatorius pagal jų sukūrimo metodiką galima suskirstyti į taisyklėmis pagrįstus ir statistinius. Statistiniais metodais veikiantys analizatoriai dar skaidomi į valdomus ir nevaldomus (angl. atitinkamai *supervised* ir *unsupervised*).

3.1.1. Taisyklėmis pagrįstas metodas

Praeito amžiaus antrojoje pusėje pradėtos kurti sistemos, skirtos informacijai iš teksto išgauti. Norint surinkti duomenis apie kokį nors asmenį ar reiškinį, svarbu rasti visas vietas tekste, kuriose jis buvo minimas. Nors anglų kalba turi labai nedaug galūnių, tačiau, ieškant visų žodžio formų, reikės pateikti keletą paieškos variantų, pvz.: *connect*, *connected*, *connecting*, *connection*, *connections*. Patogiau būtų, nurodžius vien šaknį, gauti visus to žodžio pavartojimo atvejus tekste. Taigi, svarbu nustatyti bendrą visoms formoms žodžio dalį. Martinas Porteris (Martin Porter) aprašė problemas, iškilusias nustatant žodyje ribas tarp morfemų. Naudotas priesagų atskyrimo nuo šaknies metodas, kai žodis pradedamas analizuoti nuo jo pabaigos. Tačiau greitai pastebėta, kad priesagų sąrašo nepakanka, norint 100 proc. tikslumu nustatyti šaknį. Priesaga *-er* žodžiui *sander* nustatoma teisingai, tačiau žodyje *wander* kaip priesaga *-er* traktuojama jau šaknies dalis, taip pat ir raidžių seka *ing* žodyje *reading* yra priesaga, o žodyje *sing* – šaknies dalis (Porter 1980: 313).

Tada išbandyta kita metodika – žodžio analizę atlikti nuo jo pradžios. Šaknies buvo ieškoma taip: statistiniu metodu apdorojant tekstynus, tikrinama kiekviena raidė nuo žodžio pradžios ir žiūrima, kelios skirtingos raidės gali būti iškart po jos į dešinę.

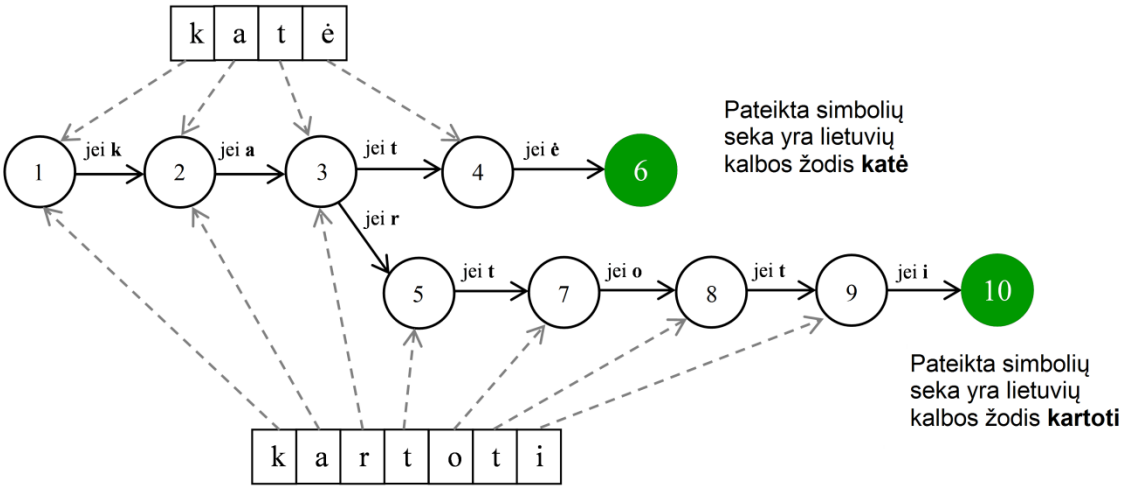
Kuo daugiau toje pozicijoje buvo surandama skirtingų raidžių, tuo didesnė tikimybė, kad ši pozicija yra jau kitos morfemos pradžia. Pavyzdžiui, anglų kalboje po raidžių rinkinio *jum* galimos tik dvi raidės: *-p* (*jump*) ir *-b* (*jumble*). O po raidžių rinkinio *jump* galimi jau penki gretimos raidės variantai: *-s* (*jumps*), *-e* (*jumped*, *jumper*), *-i* (*jumping*), *-y* (*jumpy*) ir nulinis atvejis (*jump*). Todėl laikoma, kad ši vieta ir yra riba tarp morfemų, t. y. kad pirma morfema sudaryta iš keturių raidžių – *jump*. Šio metodo trūkumas yra tas, kad jis negali atskirti, dėl kokios priežasties labai padidėja raidžių įvairovė tam tikroje pozicijoje: ar dėl morfemų ribų, ar dėl fonologinių junginių. 50 000 žodžių apimančiame anglų kalbos tekstyne po pirmosios žodžio raidės *d-* buvo rastos devynios skirtingos raidės, po dviejų raidžių junginio *de-* nustatyta jau 18 skirtingų raidžių (*dead*, *demon*, *deep*, *Delhi* ir t. t.), o po pirmų trijų raidžių *dec-* (*decided*) tegali būti tik šešios skirtingos raidės. Taigi, morfemos pabaiga klaidingai fiksuojama po antros raidės (Goldsmith 2000: 2). Atliekant žodžių morfologinę analizę vien jų suskirstymo į morfemas neužtenka. Analizuojant žodžius *jumps*, *jumping* ir nustačius, kad *jump* yra šaknis, o *-s* ir *-ing* – priesagos, dar reikia duomenų apie tai, kad *s* reiškia veiksmažodžio esamojo laiko vienaskaitos trečiąjį asmenį ir kt.

Anglų kalbos tekstynų analizei skirtose programinėse priemonėse nenumatytas kaitomų žodžių formų apdorojimas, todėl to paties kamieno variantai traktuojami kaip atskiri žodžiai (Paikens 2007: 236). Tačiau sintetinėms kalboms (tokia yra ir lietuvių kalba) toks metodas netinka, nes reikia sukaupti labai didelės apimties išteklius. Išeities buvo bandoma ieškoti įvairiais būdais. Agliutinacinių kalbų (plačiau žr. Terminų žodynėlyje), konkrečiai, suomių kalbos, morfologinei analizei atlikti buvo sukurtas dviejų lygmenų modelis (angl. *two-level model*), kuris remiasi nuostata, kad egzistuoja dviejų tipų informacija apie žodį – jo forma pavartota tekste ir pradinė forma, t. y. lema (vardažodžiams tai – vardininko linksnis, lietuvių kalbos veiksmažodžiams tai – bendratis, lotynų kalbos veiksmažodžiams – esamojo laiko vienaskaitos pirmasis asmuo ir pan.). Žodyne skiriamas tik vienas įrašas net ir tiems žodžiams, kurių šaknis kinta kaitymo metu (Koskenniemi 1983: 683), pvz., suomių kalboje daugiskaitos balsė *i*, atsidūrusi tarp dviejų balsių, virsta *j* (lietuvių kalboje panašūs atvejai galėtų būti *žaltys–žalčio*, *status–stačiam*). Dviejų lygmenų taisyklės (angl. *two-level rules*) naudojamos susieti lemą su kaitybine žodžio forma. Todėl turint vieną iš jų – lemą ar žodžio formą – pasinaudojus taisyklėmis, galima gauti kitą (Koskenniemi 1983: 684). Taisyklės kompiuteryje įdiegiamos vadovaujantis baigtinio automato (angl. *finite state automaton*) principu (Koskenniemi 1983: 683).

3.1.1.1. Baigtinis automatas

Baigtinis automatas – tai matematinis modelis, skirtas tiek fiziniams objektams, tiek abstraktiems reiškiniams pavaizduoti. Plačiausiai baigtiniai automatai naudojami projektuojant kompiuterius. Viena iš jų taikymo sričių yra ir kalbų analizė. Pagrindinė baigtinio automato savybė yra ta, kad jis turi tam tikrą baigtinį būsenų skaičių, dėl to ir vadinamas baigtiniu. Būsenos vaizduojamos grafo viršūnėmis. Baigtinis automatas gali nustatyti, ar nagrinėjama simbolių seka priklauso tam tikrai kalbai, ar ne (Dagienė, Grigas 2007: 36). Kitaip sakant, baigtinis automatas gali funkcionuoti kaip kalbos atpažinimo įtaisas. Jis pradeda darbą nuo pradinės būsenos ir po vieną simbolį skaito pateiktą simbolių seką. Perskaitęs kiekvieną simbolį, baigtinis automatas įvertina, koks tai simbolis, ir atsižvelgia į būseną, kurioje jis yra to simbolio skaitymo metu. Pagal šiuos du parametrus ir sprendžia, į kurią kitą būseną jam reikia pereiti. Perskaitęs paskutinį sekos simbolį, baigtinis automatas turi atsidurti galinėje būsenoje, jei darbas baigtas sėkmingai, t. y. jei perskaityta simbolių seka buvo atpažinta kaip leistinas toje kalboje raidžių rinkinys. Jeigu perskaitytas paskutinis simbolis neatveda į galinę būseną, rodančią sėkmingą atpažinimą, vadinasi, pateikta simbolių seka nepriklauso tai kalbai. Taigi, turėdami bet kokią simbolių seką, galime patikrinti, ar ji priklauso kalbai, kurią apibrėžia mūsų sukurtas baigtinis automatas. Labai paprasto baigtinio automato pavyzdys pateiktas 40 pav. Šis automatas atpažįsta tik du lietuvių kalbos žodžius: *katė* ir *kartoti*, t. y. tik šie raidžių rinkiniai gali atvesti į galines automato būsenas – šeštąją ir dešimtąją. Pateikus raidžių rinkinį *karbgd* jis jau nebus atpažintas kaip galimas lietuvių kalbos žodis, nes darbas sustos penktojoje būsenoje todėl, kad iš jos neišeina linija į jokią kitą viršūnę, jei perskaitytas simbolis yra raidė *b*. Perskaitęs paskutinį simbolį, baigtinis automatas tebebus penktojoje būsenoje, t. y. jis nepereis nė į vieną iš galinių būsenų (šeštąją arba dešimtąją). Vadinasi, darbo pabaiga yra nesėkminga, ir pranešimas bus toks: *Pateiktos simbolių sekos nepavyko atpažinti kaip lietuvių kalbos žodžio*. Šis pavyzdys labai gerai parodo, kad baigtinio automato perėjimas į kitą būseną priklauso nuo perskaityto simbolio ir nuo būsenos, kurioje jis tuo metu yra: jei simbolis *t* perskaitomas, kai automatas yra trečiojoje būsenoje, jis pereina į ketvirtąją būseną. Jei tas pats simbolis *t* perskaitomas esant automatui penktojoje būsenoje, jis pereina į septintąją būseną.

Baigtinio automato principas buvo naudotas ir kuriant pirmąjį lietuvių kalbos morfologinį analizatorių. Programinės įrangos autorius – V. Zinkevičius (Zinkevičius 2000: 252).



40 pav. Baigtinis automatas, atpažįstantis du lietuvių kalbos žodžius: *katė* ir *kartoti*

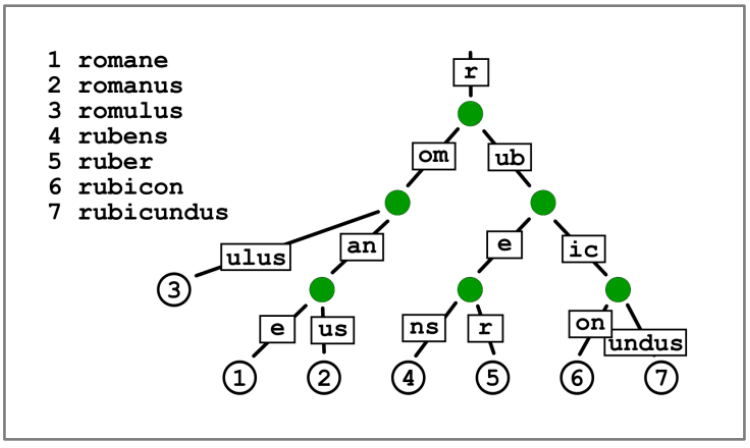
3.1.1.2. Lietuvių kalbos *Lemuoklis*

Atliekant lietuvių kalbos morfologinę analizę taikomas suomių kalbai sukurtas dviejų lygmenų modelis (Koskenniemi 1983: 684), t. y. traktuojama, kad žodis turi kaitybinę formą ir žodynuose vartojamą antraštinę formą – lemą (Zinkevičius 2000: 245). Suomių kalba priklauso agliutinacinių kalbų grupei. Pagrindinis skirtumas tarp agliutinacinių ir fleksinių kalbų (tokia yra ir lietuvių kalba) išryškėja bandant skaidyti žodžius į pastovią ir kintamą dalis. Agliutinacinėse kalbose žodį sudaro šaknis ir afiksai, kurių kiekis įvairiose to paties žodžio formose yra nevienodas, pvz., skirtingi to paties žodžio linksniai gali turėti skirtingą priesagų kiekį. Fleksinėse kalbose žodis traktuojamas kaip kamieno ir galūnės junginys, nes kaitymo metu kinta tik jo galūnė, o šaknis su priesaga (ar priesagomis) lieka pastovios (Aleksa 2006: 3). Dar viena fleksinių kalbų savybė, kuri skiria jas nuo agliutinacinių kalbų, – tai dažnai pasitaikanti atsitiktinė morfemų kaita. Bandant ją aprašyti taisyklėmis, kiekvienai paradigmai tenka sukurti naują morfologinę klasę. Tai atlikus, gaunamas labai didelis jų skaičius: čekų kalbai – apie 1 500, rusų kalbai – apie 1 000 morfologinių klasių (Gelbukh, Sidorov 2003: 216). Lietuvių kalbos *Lemuokliui* buvo sudaryta 600 morfologinių klasių.

Latvių kalbininkai pabrėžia daugiareikšmiškumo problemą. Šiuo metu latvių kalbos morfologinis analizatorius daugiau nei pusei (50–55 proc.) žodžių pateikia kelis

galimus variantus, vidutiniškai 3–4 alternatyvas (Paikens, Rituma, Pretkalniņa 2013: 271). Nurodoma, kad ypač sunku nustatyti įvardžio tipą, nes daugeliu atvejų tai priklauso nuo žodžio vietos sakinyje, pvz., įvardis *kas* gali būti nežymimasis, klausiamasis arba santykinis (Levane, Spektors 2000: 1097). Lietuvių kalbos *Lemuoklis* taip pat daugeliui žodžių pateikia kelis galimus pradinės formos variantus. Daugiareikšmiškumą panaikinti gali tiktai kontekstas ar sintaksinė sakinio analizė.

Lemuoklio kalbinių duomenų bazę sudaro šeši kompiuteriniai žodynai: GF (Gramatinės formos), T (Tikriniai daiktavardžiai), S (Santrumpos ir akronimai) ir kt. Visi žodynai suformuoti kaip medžio formos duomenų struktūros. Svarbiausias žodynas yra GF. Jis naudojamas, kai reikia atpažinti lietuvių kalbos žodžių kaitybines formas ir pateikti jų morfologinius duomenis. Šis žodynas sudarytas iš dviejų komponentų: žodžių šaknų sąrašo ir kompiuterinės morfologijos, t. y. skaitmeninių lietuvių kalbos kaitybos modelių (Zinkevičius 2000: 252). Šaknims duomenų bazėje pavaizduoti naudojamas raidžių medis (angl. *radix tree* arba *radix trie* iš „information re**Trie**val“). Tai suspausto medžio formatas, kai į vieną sujungiamos šalia esančios viršūnės, jei iš jų išeina tik po vieną lanką. Šis metodas buvo sukurtas paieškai tekstuose atlikti[60] (38 interneto nuoroda[61]). 41 pav. parodytas raidžių medžio pavyzdys. Kompiuterinė informacijos paieška ir išrinkimas tokiose struktūrose vyksta labai greitai. Medžių šakas sudaro raidžių sekos, o ieškoma informacija įrašyta jo lapuose (Zinkevičius 2000: 256).



41 pav. Raidžių medžio (angl. *radix tree*) pavyzdys (39 interneto nuoroda[62])

[60] **P**ractical **A**lgorithm **T**o **R**etrieve **I**nformation **C**oded **I**n **A**lphanumeric (38 interneto nuoroda).

[61] Prieiga internete: https://de.wikipedia.org/wiki/Patricia-Trie [žiūrėta 2022-11-22].

[62] Prieiga internete: http://en.wikipedia.org/wiki/Radix_tree [žiūrėta 2022-11-22].

Lemuoklyje prie kiekvienos šaknies, esančios sąraše, nurodoma jos morfologinė klasė. Kompiuterinėje morfologijoje tai yra taisyklių rinkinys, nusakantis, kaip ir kokius afiksus galima jungti prie tokio tipo šaknų ir kokias gramatines reikšmes įgyja padarytos žodžių formos. Atlikdamas žodžio analizę, *Lemuoklis* jį traktuoja kaip raidžių seką ir bando ją suskaidyti pagal schemą (42 pav.) į įvairius variantus (Zinkevičius 2000: 252).

42 pav. Žodžio skaidymo schema *Lemuoklyje* (parengta pagal Zinkevičius 2000: 252)

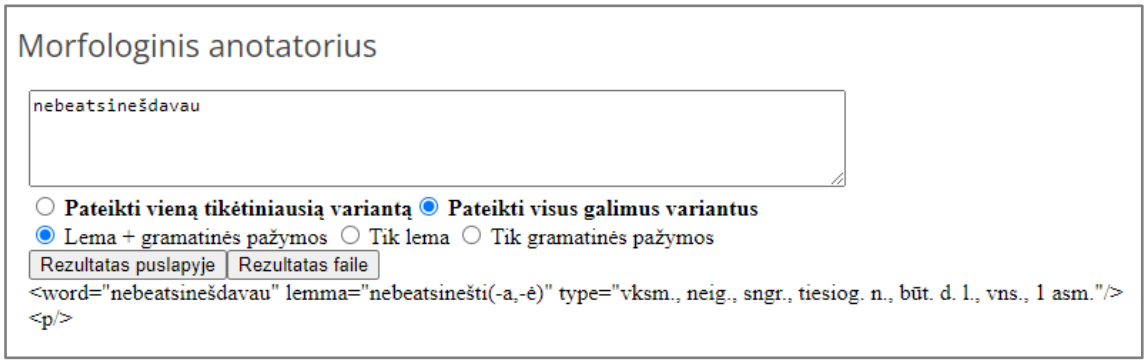
Jei nors vieno varianto šakniai randamas atitikmuo šaknų sąraše, tada kompiuterinėje morfologijoje ieškoma tai šakniai galimų prefiksų ir postfiksų. Jei ir jie randami, žodis laikomas atpažintu ir pateikiami jo morfologiniai duomenys.

Kuriant morfologinius analizatorius taisyklėmis pagrįstu metodu, reikia įdėti labai daug darbo ir jis turi būti atliekamas labai aukštos kvalifikacijos specialistų. Tačiau naudojant šį metodą nereikia jokių papildomų išteklių, tokių kaip anotuoti tekstynai. Tai gali būti priežastis, kodėl šiuo metu taisyklėmis pagrįsti analizatoriai dar paplitę[63] (Boizou, Kapočiūtė-Dzikienė, Rimkutė 2018: 19). *Lemuoklis* ir šiuo metu naudojamas VDU tiekiamoje anotavimo paslaugoje *Morfologinis anotatorius* (40 interneto nuoroda[64]). Jo trūkumai būtų tokie: neatpažįsta žodžių su sutrumpėjusiomis galūnėmis, pvz., *šnekamojoj kalboj*, kartais kaip lemą nurodo perteklinius, lietuvių kalboje neegzistuojančius žodžius, pvz.: *blizgėjas*, *žibėjas*. Tokių žodžių analizės rezultatuose neatsiranda, jei pasirenkamas režimas, pateikiantis tik vieną, labiausiai tikėtiną variantą. Tačiau šiuo atveju prarandami ir labai dažnai vartojami žodžiai, pvz., žodžio formai *laikai* kaip lema lieka tik daiktavardžio vardininkas, nors kažin ar veiksmažodį *laikyti* galima priskirti prie labai retai vartojamų. Žodžiui *stovi* tokiame režime nurodomas tik daiktavardžio *stovis* šauksmininkas ir vargu ar jis yra labiau tikėtinas nei veiksmažodžio *stovėti* esamojo laiko trečiasis asmuo. Šių trūkumų neturi morfologiškai anotuotas *Dabartinės lietuvių*

[63] "Probably due to this reasons this approach is still the most widely spread" (Boizou, Kapočiūtė-Dzikienė, Rimkutė 2018: 19).

[64] Prieiga internete: https://klc.vdu.lt/anotatorius/ [žiūrėta 2022-11-22].

kalbos tekstynas (25 interneto nuoroda[65]), tačiau kai kurių rečiau vartojamų formų jame iš viso nėra, pvz., *nebeatsinešdavau* (41 interneto nuoroda[66]). *Morfologinis anotatorius* šiam žodžiui pateikia visiškai tikslią morfologinę informaciją (43 pav.).



43 pav. Morfologinio anotatoriaus (40 interneto nuoroda[67]) pateikiama informacija žodžiui *nebeatsinešdavau*

3.1.1.3. Morfologinis analizatorius *semantika.lt*

2011–2015 m. buvo sukurtas dar vienas taisyklių pagrindu veikiantis lietuvių kalbos morfologinis analizatorius, tik jau nenaudojantis baigtinio automato principo gramatikos taisyklėms aprašyti. Tai atvirojo kodo *Hunspell* platformoje parengtas lietuvių kalbos morfologinis analizatorius. Jis buvo naudojamas VDU *Lietuvių kalbos sintaksinės ir semantinės analizės informacinėje sistemoje*, tinklalapyje *semantika.lt*. *Hunspell* buvo sukurtas vengrų kalbai (iš čia ir pavadinimo pirmasis dėmuo *Hun-* pagal Vengrijos valstybės pavadinimą anglų kalba) – tai vengrų kalbos rašybos klaidų tikrintuvas. Ši kalba priklauso agliutinacinių kalbų grupei, tačiau dėl programinio kodo universalumo jį galima taikyti ir kito tipo kalboms. Pritaikant analizatorių lietuvių kalbai buvo sudaryta apie 18 000 taisyklių. Vienos tokios taisyklės pavyzdys (Dadurkevičius 2017: 3) pateiktas 44 pav.

```
SFX  85  čias  ty  [^š]čias  is:Masc_Sg_Voc
```

44 pav. *Hunspell* platformoje parengto lietuvių kalbos morfologinio analizatoriaus taisyklė (parengta pagal Dadurkevičius 2017: 3)

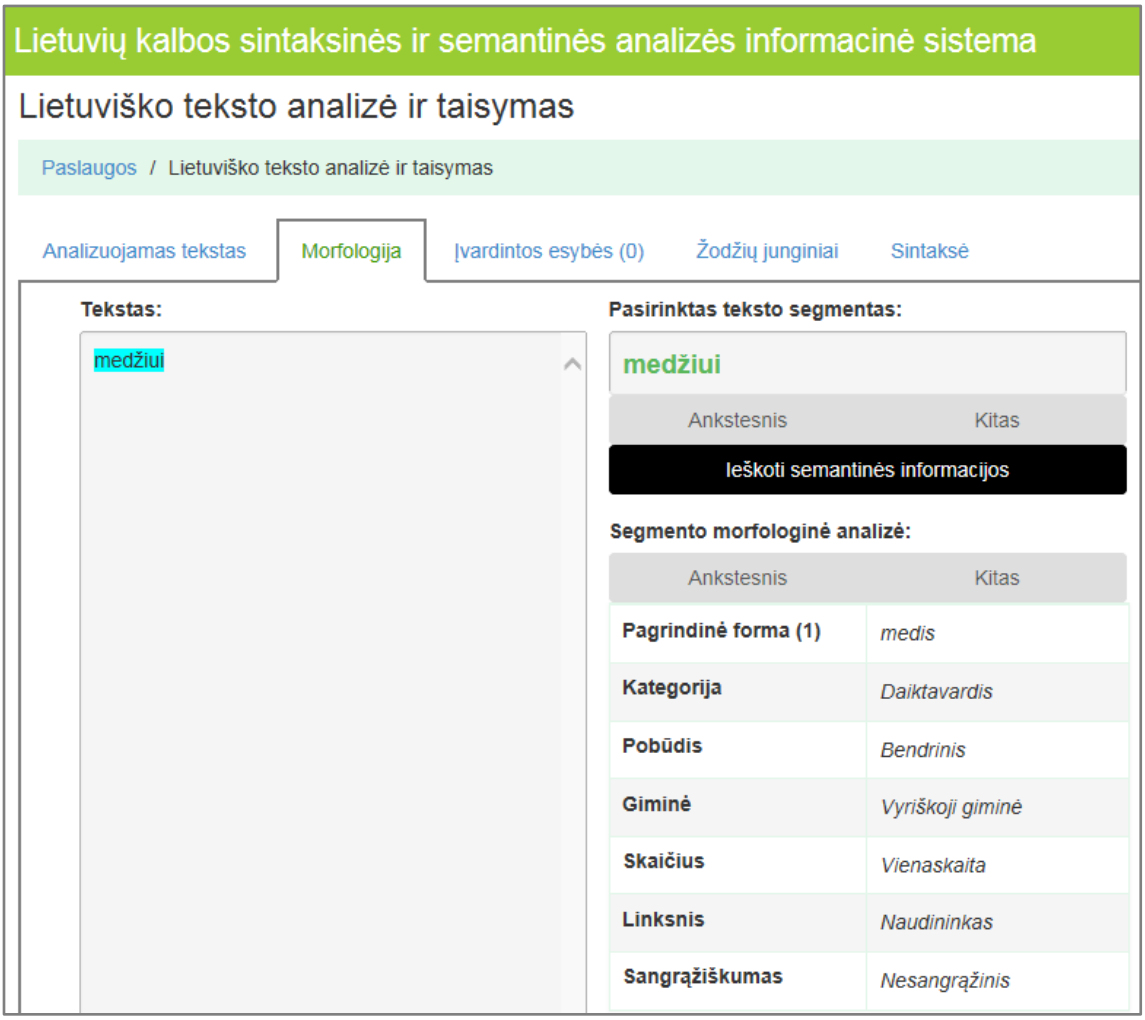
[65] Prieiga internete: https://klc.vdu.lt/matas-morfologiskai-anotuotas-tekstynas/ [žiūrėta 2022-11-22].
[66] Prieiga internete: http://corpus.vdu.lt/lt/?word=nebeatsine%C5%A1davau [žiūrėta 2022-11-22].
[67] Prieiga internete: https://klc.vdu.lt/anotatorius/ [žiūrėta 2022-11-22].

Šioje taisyklėje raidžių junginys *SFX* reiškia, kad keičiama žodžio pabaiga (*PFX* rodytų, kad keičiama žodžio pradžia). Skaičius *85* yra taisyklių grupės, kurią gali sudaryti viena arba kelios taisyklės, taikomos kartu, numeris. 44 pav. taisyklė nurodo, kad žodžio pabaigą *-čias* galima keisti į *-ty*, pvz., *svečias – svety*. Tačiau šioje taisyklėje yra įvestas apribojimas, kad jos taikyti negalima, jei prieš *-čias* yra raidė *-š-*, t. y. žodis baigiasi raidžių junginiu *-ščias*, pvz., *vaikiščias*. Morfologinis pažymėjimas *Masc_Sg_Voc* reiškia, kad po pakeitimo gaunama vyriškosios giminės vienaskaitos šauksmininko forma (Dadurkevičius 2017: 3). Atliekant analizę taisyklių metodu pateikiami visi galimi žodžio morfologinių požymių variantai. Siekiant panaikinti daugiareikšmiškumą, analizatoriuje *semantika.lt* buvo naudojami statistiniai metodai.

Morfologinė analizė yra pirmas automatinio kalbos apdorojimo etapas. Pagrindinis jos tikslas – paruošti sakinį tolesnei sintaksinei ar semantinei analizei. Todėl labai svarbu, kad ji būtų atlikta kuo tiksliau, nes morfologijos lygmenyje padarytos klaidos trukdo atliekant automatinę kitų lygmenų kalbos analizę (Bielinskienė, Boizou, Rimkutė 2017: 2). Buvo atlikti abiejų lietuvių kalbos morfologinių analizatorių – *Lemuoklis* (V. Zinkevičiaus morfologinės analizės ir sintezės programinė įranga papildyta vienareikšminimo funkcija, žr. 2.2.3 poskyrį) ir *semantika.lt* – tyrimai, siekiant išsiaiškinti jų pateikiamų rezultatų tikslumą. Palyginus gautus duomenis, buvo nustatyta, kad skirtumai labai nedideli. Bendras tikslumo procentas aukštesnis *semantika.lt*, tačiau atskiriems tekstams, pvz., moksliniams, geresnius rezultatus pateikė *Lemuoklis*. Nevienodus rezultatus atskirose srityse galėjo lemti struktūriniai skirtumai: *Lemuoklis* turi mažiau lemų nei *semantika.lt*, tačiau jame naudojamas sintezės metodas. Tyrimo metu buvo analizuojami rezultatai pagal 32 morfologinius požymius iš aštuonių kategorijų: linksnis, giminė, skaičius, laipsnis, laikas, nuosaka, asmuo, rūšis. *Lemuoklio* rezultatai geresni buvo pagal 12 požymių, *semantika.lt* – pagal 11. Vardininko linksnį geriau atpažino *Lemuoklis*. Daugeliu atvejų (apie 5 proc.) *semantika.lt* neteisingai žymėjo vardininką – pateikė jį kaip šauksmininką. Neteisingai sužymėtų šauksmininkų buvo daugiau nei nurodytų teisingai. Tai turėjo įtakos bendram vardininko rezultatui. *Lemuoklis* daugiau klaidų padarė įnagininko atveju, bet jis geriau atpažino dalyvius. Tyrimo autorių išvada: geresni *Lemuoklio* rezultatai analizuojant mokslinius tekstus gauti todėl, kad juose buvo didesnis dalyvių procentas nei kitų sričių tekstuose (Boizou, Kapočiūtė-Dzikienė, Rimkutė 2018: 22). Tikėtina, kad tuo pačiu metodu (pagrįstu taisyklėmis) veikiantys analizatoriai pateikia panašaus tikslumo rezultatus. Ateityje planuojami testai

naudojant kitų tipų tekstus: forumo pranešimus, interneto komentarus ir kt. (Kapočiūtė-Dzikienė, Rimkutė, Boizou 2017: 54).

Nuo 2016 m. balandžio 25 d. iki 2020 m. vasario 21 d. *Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema* buvo laisvai prieinama internete (42 interneto nuoroda[68]). 45 pav. parodyta žodžio *medžiui* analizė (43 interneto nuoroda[69]).

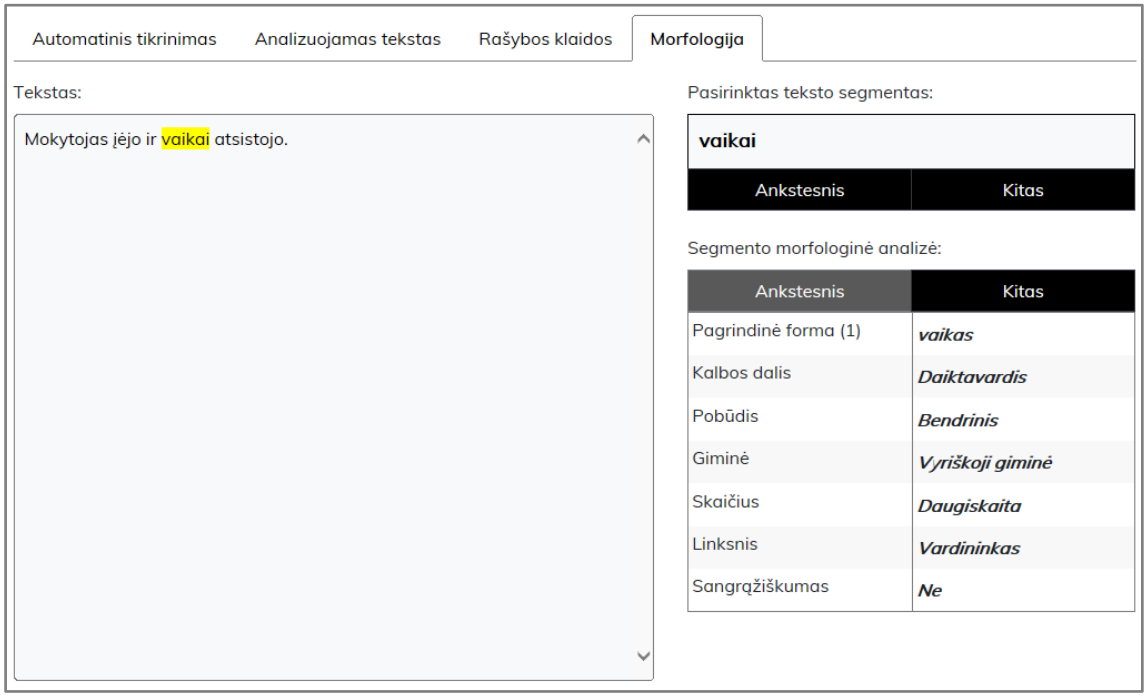


45 pav. Žodžio *medžiui* morfologiniai duomenys (43 interneto nuoroda)

[68] Prieiga internete: https://web.archive.org/web/20200221090527/http://www.semantika.lt:80/SyntaticAndSemanticAnalysis/Analysis [žiūrėta 2022-11-22].

[69] Prieiga internete: http://www.semantika.lt/SyntaticAndSemanticAnalysis/Analysis [žiūrėta 2016-10-26].

2020 m. parengta nauja tinklalapio *semantika.lt* versija (44 interneto nuoroda[70]). 46 pav. pateikta sakinio *Mokytojas įėjo ir vaikai atsistojo* analizė, tiksliau, žodžio *vaikai* morfologiniai duomenys. Visų šio sakinio žodžių morfologinę analizę galima pamatyti 6 priede. Daugiareikšmiams žodžiams pateikiami visi galimi variantai.



46 pav. Sakinio *Mokytojas įėjo ir vaikai atsistojo* žodžio *vaikai* analizė (44 interneto nuoroda)

3.1.2. Statistiniai metodai

Automatinis mokymasis būna dviejų pagrindinių tipų: valdomas (angl. *supervised learning*) ir nevaldomas (angl. *unsupervised learning*).

3.1.2.1. Valdomas mokymasis

Valdomo mokymosi atveju žmogus pateikia kompiuteriui apmokymo duomenų rinkinį – pradinių duomenų ir teisingų rezultatų pavyzdžius. Kompiuteris iš jų išgautą informaciją (modelį) įsirašo į savo atmintį. Gavęs naujus duomenis, jis analizuoja, kas yra jo atmintyje, ir pagal turimą informaciją bando paruošti rezultatą naujai gautiems duomenims.

[70] Prieiga internete: https://www.semantika.lt/Analysis/TextAnalysis [žiūrėta 2022-11-22].

Olandų kalbos morfologinis analizatorius. Tokiu metodu yra sukurtas olandų kalbos morfologinis analizatorius. 47 pav. parodyta žodžio *abnormaliteiten* analizė. Žodis padalijamas į segmentus, kuriuose aplink kiekvieną raidę paliekama po penkis simbolius iš kairės ir iš dešinės. Iš viso segmente būna ne daugiau kaip 11 simbolių, o pačių segmentų – tiek, kiek žodyje raidžių. Pirmasis segmentas, remiantis kompiuterio atmintyje įrašyta informacija, priskiriamas klasei *A+Da*. Ši žyma parodo, kad pirmoji morfema *abnormal* yra būdvardis (*A*) ir kad tai dar ne visas žodis, kad dar yra jo pabaiga, kuri šiame segmente nutrinta (*+Da*). Peržiūrint devintąjį segmentą, fiksuojama, kad antroji morfema *-iteit-* priklauso klasei *N_A**. Tai reiškia, kad, prijungus šią morfemą prie būdvardžio (*A**) iš dešinės pusės, gaunamas daiktavardis *N*. Keturioliktame segmente iš atminties paimama informacija, kad galūnė *-en* rodo daugiskaitą, todėl trečioji morfema pažymima simboliu *m* (Bosch, Daelemans 1999: 286). Gautas žodžio analizės rezultatas pateiktas 48 pav.

instance number	left context					focus letter	right context					TASK
1	-	-	-	-	-	a	b	n	o	r	m	A+Da
2	-	-	-	-	a	b	n	o	r	m	a	0
3	-	-	-	a	b	n	o	r	m	a	l	0
4	-	-	a	b	n	o	r	m	a	l	i	0
5	-	a	b	n	o	r	m	a	l	i	t	0
6	a	b	n	o	r	m	a	l	i	t	e	0
7	b	n	o	r	m	a	l	i	t	e	i	0
8	n	o	r	m	a	l	i	t	e	i	t	0
9	o	r	m	a	l	i	t	e	i	t	e	N_A*
10	r	m	a	l	i	t	e	i	t	e	n	0
11	m	a	l	i	t	e	i	t	e	n	-	0
12	a	l	i	t	e	i	t	e	n	-	-	0
13	l	i	t	e	i	t	e	n	-	-	-	0
14	i	t	e	i	t	e	n	-	-	-	-	m
15	t	e	i	t	e	n	-	-	-	-	-	0

47 pav. Olandų kalbos žodžio *abnormaliteiten* morfologinė analizė
(parengta pagal Bosch, Daelemans 1999: 288)

$$[\text{abnormal}]_{A}[\text{iteit}]_{N\text{-}A*}[\text{en}]_{m}$$

48 pav. Olandų kalbos žodžio *abnormaliteiten* morfologinės analizės rezultatas
(parengta pagal Bosch, Daelemans 1999: 288)

Tokius analizatorius nesunku patobulinti pateikus jiems daugiau anotuotų tekstų (Boizou, Kapočiūtė-Dzikienė, Rimkutė 2018: 19).

Lietuvių kalbos morfologinis analizatorius. Morfologinę lietuvių kabos analizę atlieka statistinis *UDPipe* konvejeris (angl. *pipeline*), naudojantis giliųjų neuroninių tinklų metodus (45 interneto nuoroda[71]). Jis apima apie 50 kalbų. Gali atlikti sakinio segmentavimą ir morfologinę bei sintaksinę jo analizę (Straka, Strakova 2017: 88). Automatinio mokymosi metu analizatoriui turi būti pateiktas nedidelis anotuotas tekstynas CoNLL-U fotmatu. Pagal gautus duomenis sudaromos taisyklės, kaip turi būti generuojama žodžio lema: atmetus kai kuriuos priešdėlius ir priesagas, pridedamos kitos priesagos (Straka, Strakova 2017: 91) Pavyzdžiui, vokiečių kalbos dalyvis padaromas su priešdėliu *ge-* ir priesaga *-t* (pvz., *gestellt*), taigi, ieškant lemos, reikia juos atmesti ir pridėti bendraties priesagą *-en* (*stellen*). Lietuvių kalbos apmokymo duomenims buvo naudotas sintaksiškai anotuotas lietuvių kalbos tekstynas ALKSNIS.

3.1.1.3 poskyryje buvo aptarta sakinio *Mokytojas įėjo ir vaikai atsistojo* analizė (46 pav.) taisyklėmis pagrįstu metodu. *UDPipe*, t. y. statistiniais metodais, atliktos šio sakinio analizės fragmentas pavaizduotas 49 pav. Žodžiui *įėjo* neteisingai nustatyta lema – *įėti*.

A Output Text | Show Table

Save Output File

Id	Form	Lemma	UPosTag	XPosTag	Feats
# text = Mokytojas įėjo ir vaikai atsistojo.					
1	Mokytojas	mokytojas	NOUN	dkt.vyr.vns.V.	Case=Nom|Gender=Masc|Number=Sing
2	įėjo	įėti	VERB	vksm.asm.tiesiog.būt-k.vns.3.	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Past|VerbForm=Fin
3	ir	ir	CCONJ	jng.	_
4	vaikai	vaikas	NOUN	dkt.vyr.dgs.V.	Case=Nom|Gender=Masc|Number=Plur
5	atsistojo	atsistoti	VERB	vksm.asm.sngr.tiesiog.būt-k.dgs.3.	Aspect=Perf|Mood=Ind|Number=Plur|Person=3|Polarity=Pos|Reflex=Yes|Tense=Past|VerbForm=Fin
6	.	.	PUNCT	skyr.	_

49 pav. Sakinio *Mokytojas įėjo ir vaikai atsistojo* analizės, atliktos analizatoriumi *UDPipe*, fragmentas (45 interneto nuoroda)

[71] Prieiga internete: https://lindat.mff.cuni.cz/services/udpipe/run.php [žiūrėta 2022-11-22].

Tikėtina, kad *UDPipe* apmokymo duomenyse nebuvo žodžio *įeiti* bei jo formų. Analizatorius pagal susikurtas lemos nustatymo taisykles sugeneravo šio žodžio pradinę formą, matomai, remdamasis analogija su kitais žodžiais: žodžiui *kalbėjo* – atmetama galūnė *-jo*, pridedama priesaga *-ti*, gaunama *kalbėti*, žodžiui *stovėjo* – atmetama galūnė *-jo*, pridedama priesaga *-ti,* gaunama *stovėti*, taigi, ir žodžiui *įėjo* buvo atmesta galūnė *-jo*, pridėta priesaga *-ti* ir gauta *įėti*.

Šios klaidos taisyklėmis pagrįstas, *Hunspell* platformoje veikiantis morfologinis analizatorius *semantika.lt* nepadarė (6 priedas).

Kol lietuvių kalbos sakinio struktūra panaši į anglų kalbos, t. y. jame pavartota tipiška anglų kalbos žodžių tvarka (veiksnys – tarinys – netiesioginis papildinys – tiesioginis papildinys, pvz., *I give her an apple*), morfologiniai analizatoriai dirba gerai. Sakinys *Mama padovanojo man mažą šuniuką* išnagrinėjamas be klaidų naudojant abiejų tipų metodikas, t. y. tiek taisyklių pagrindu veikiančiu analizatoriumi, tiek statistiniais metodais. Tačiau, jei sakinyje yra lietuvių kalbai būdinga, laisva, neangliška žodžių tvarka, padaroma nemažai klaidų. Atrodytų, toks nesudėtingas sakinys *Lauke sninga* išanalizuojamas neteisingai: žodžiui *sninga* kaip lema nustatomas būdvardžio vardininkas *sningas*, o kaip tekste pavartota forma nurodoma jo bevardė giminė (50 pav.). Šios klaidos priežastis gali būti panaši kaip ir nustatant žodžio *įėjo* lemą. Matomai, pagal analogiją su *lauke gera* – šiuo atveju tai tikrai būdvardžio bevardė giminė, o lema gaunama atmetus galūnę *-a* ir pridėjus vardininko galūnę *-as* (*geras*) – buvo gauta ir lema *sningas*.

Save Output File

Id	Form	Lemma	UPosTag	XPosTag	Feats
# text = Lauke sninga.					
1	Lauke	laukas	NOUN	dkt.vyr.vns.Vt.	Case=Loc|Gender=Masc|Number=Sing
2	sninga	sningas	ADJ	bdv.nelygin.bev.	Definite=Ind|Degree=Pos|Gender=Neut
3	.	.	PUNCT	skyr.	_

50 pav. Sakinio *Lauke sninga* analizės, atliktos analizatoriumi *UDPipe*, fragmentas (45 interneto nuoroda)

Iš poezijos paimtas sakinys *Te pasakų griūva skliautai gintariniai* morfologiškai išanalizuojamas taip pat su klaida: dalelytei *te* nustatoma, kad tai įvardis, vyriškoji giminė, daugiskaitos vardininkas.

Dar vienas sakinys, kuriam būdinga neangliška žodžių tvarka, – *Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus* – išanalizuotas su dviem klaidomis: žodžiui *Liūdną* priskiriamas tikrinio daiktavardžio požymis ir žodis *šuo* traktuojamas kaip įvardžio įnagininkas, kurio pradinė forma (lema) yra *šas* (51 pav.).

A Output Text | Show Table

Save Output File

Id	Form	Lemma	UPosTag	XPosTag	Feats
# text = Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus.					
1	Liūdną	Liūdnas	PROPN	dkt.tikr.mot.vns.G.	Case=Acc|Gender=Fem|Number=Sing
2	jis	jis	PRON	įv.vyr.vns.V.	Case=Nom|Definite=Ind|Gender=Masc|Number=Sing|Person=3|PronType=Prs
3	mums	aš	PRON	įv.dgs.N.	Case=Dat|Definite=Ind|Number=Plur|Person=1|PronType=Prs
4	pranešė	pranešti	VERB	vksm.asm.tiesiog.būt-k.vns.3.	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Polarity=Pos|Tense=Past|VerbForm=Fin
5	žinią	žinia	NOUN	dkt.mot.vns.G.	Case=Acc|Gender=Fem|Number=Sing
6	:	:	PUNCT	skyr.	_
7	šuo	šas	DET	įv.vyr.vns.Įn.	Case=Ins|Definite=Ind|Gender=Masc|Number=Sing|PronType=Dem
8	nebegrįžo	nebegrįžti	VERB	vksm.asm.neig.tiesiog.būt-k.vns.3.	Aspect=Perf|Mood=Ind|Number=Sing|Person=3|Polarity=Neg|Tense=Past|VerbForm=Fin
9	į	į	ADP	prl.G.	AdpType=Prep|Case=Acc
10	namus	namai	NOUN	dkt.vyr.dgs.G.	Case=Acc|Gender=Masc|Number=Plur
11	.	.	PUNCT	skyr.	_

51 pav. Sakinio *Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus* analizės, atliktos analizatoriumi *UDPipe* , fragmentas (45 interneto nuoroda)

Taigi, pritartina labai teisingam lietuvių kalbos kompiuterizavimo srityje dirbusio Vido Daudaravičiaus teiginiui: „Naivu manyti, kad metodai, kurie sėkmingai taikomi anglų kalbai, tinka ir kitoms kalboms“ (Daudaravičius 2012: 3). Panašių minčių išsako ir kiti autoriai. Neuroniniai tinklai naudojami įvairiose srityse – vaizdų atpažinimo; garso, sakytinės šnekos atpažinimo; automatinio vertimo, tačiau esant dideliam duomenų išsibarstymui (kas ypač būdinga kalboms) jie neduoda labai gerų rezultatų (Choi 2016: 271).

3.1.2.2. Nevaldomas mokymasis

Jau praeito amžiaus pabaigoje buvo bandoma sukurti programinę įrangą, kuri gavusi gryną, neanotuotą tekstą sugebėtų atlikti jo analizę be žmogaus įsikišimo (Sinclair 1992: 381). Tai iš dalies galima pasiekti naudojant kito tipo morfologinės analizės algoritmus, paremtus nevaldomu apmokymu (angl. *unsupervised learning*). Esminis šio metodo bruožas yra tas, kad vieninteliai įvedami duomenys yra tekstynas. Programinė įranga turi tik analizės įrankius, tačiau jokio žodyno ar morfologinių taisyklių, būdingų konkrečiai kalbai, ji negauna[72] (Goldsmith 2001: 154). Pagrindinis tikslas – „teisingai suskaidyti žodį į jo sudedamąsias dalis (morfemas), pateikiant tik pradinę, t. y. pačią bendriausią, morfologinę informaciją“[73] (Goldsmith 2001: 154). Galima sakyti, kad šio metodo pritaikymas būtų labiau pačių gramatikų rašymas, automatinis morfologijos sukūrimas. Tokias gramatikas aktualu parašyti istoriniams tekstams ar kalboms, kurias vartojančių gimtakalbių jau nebėra. Populiariai automatinio morfologijos sukūrimo eigą galima aprašyti taip: iš pradžių ieškoma žodžio kamieno kaip priešpriešos kaitybiniams afiksams, toliau bandoma surasti darybinius priešdėlius ir priesagas. Tikslas – nustatyti, kad viename tekstyne dažniausiai pasitaikančios priesagos yra *-s*, *-ing*, *-ed*; o kitame tekstyne pagrindinės priesagos yra *-e*, *-en*, *-heit*, *-ig*. Kadangi programinė įranga nėra skirta kalboms atpažinti, ji nepasakys, kad pirmasis tekstynas yra anglų kalbos, o antrasis – vokiečių. Tačiau kiekvienam žodžiui bus nurodyta, kur yra jo kamienas ir kur – afiksai. Žodis traktuojamas kaip susidedantis iš šaknies, prieš kurią gali būti priešdėliai, o po jos – priesagos.

2015 m. buvo pasiūlytas nevaldomu mokymusi pagrįstas metodas morfosintaksiniams žodynams sudaryti: turint 1 000 žodžių apimties pradinį žodyną (angl. *seed*), jo apimtį galima padidinti 100 kartų. Tačiau panaudotas šis metodas buvo tik 11-ai kalbų (Faruqui, McDonald, Soricut 2016: 6).

[72] “[...] program’s sole input is the corpus; we provide the program with the tools to analyse, but no dictionary and no morphological rules particular to any specific language” (Goldsmith 2001: 154).

[73] “[...] the goal of the program is restricted to providing the correct analysis of words into component pieces (morphemes), though with only a rudimentary categorical labeling” (Goldsmith 2001: 154).

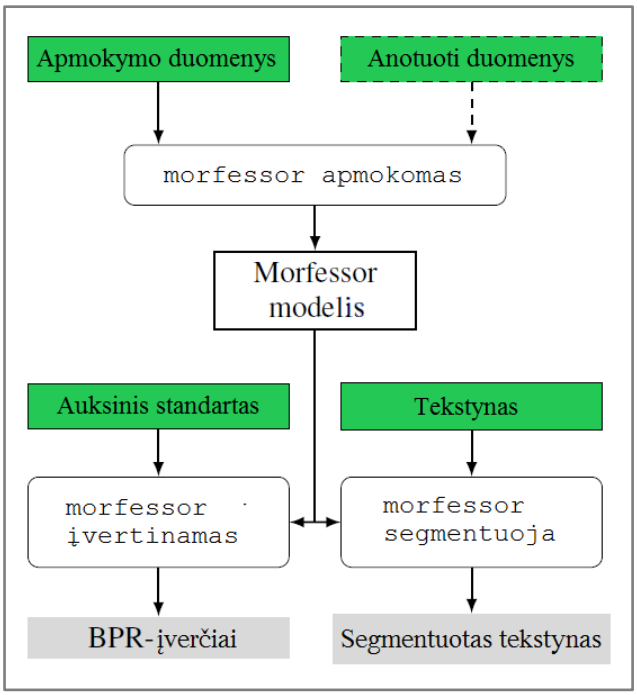
52 pav. parodyta, kokie duomenys gauti anglų ir italų kalbų žodžiams naudojantis pradinio žodyno 1 000 žodžių duomenimis (Faruqui, McDonald, Soricut 2016: 10). Padidinus jo apimtį nuo 1 000 iki 5 000 žodžių, tikslumas padidėjo apie 10 proc., tačiau, padidinus apimtį nuo 5 000 iki 10 000 žodžių, tikslumas tepadidėjo tik apie 2 proc.

	Word	Attributes
en	study (seed)	POS:Verb, VForm:Fin, Mood:Ind, Tense:Pres, Num:Sing, POS:Noun
	studied	POS:Verb, VForm:Fin, Mood:Ind, Tense:Past, VForm:Part
	taught	POS:Verb, VForm:Fin, Mood:Ind, Tense:Past, VForm:Part, Voice:Pass
it	tavola (seed)	POS:Noun, Gender:Fem, Num:Sing
	tavoli	POS:Noun, Gender:Masc, Num:Plur
	divano	POS:Noun, Gender:Masc, Num:Sing

52 pav. Anglų ir italų kalbų morfosintaksinio žodyno fragmentas (parengta pagal Faruqui, McDonald, Soricut 2016: 10)

Kita programinė įranga *Morfessor* taip pat paima tekstyną kaip duomenis ir suskaido žodžius morfemomis be žmogaus pagalbos. Tačiau čia traktuojama, kad žodis turi kintamą skaičių morfemų, ir tas skaičius iš anksto nėra žinomas. Analizuodamas kalbos duomenis, *Morfessor* pasiūlo modelį, kuriame atsispindi dėsningumai, pastebėti pateiktame žodžių formų rinkinyje. Pagrindinis tikslas – sukurti optimalų morfemų žodyną, kuriuo remiantis, būtų galima segmentuoti tekstyno žodžius (Creutz ir kt. 2005: 3). Probleminiais morfemų ribų nustatymo atvejais naudojama neapibrėžtų ribų tarp morfemų nustatymo funkcija, pvz., angliškas žodis *tyrannizes* gali būti skaidomas dvejopai: *tyrann-ize-s* ir *tyrann-iz-es*, nes žodyje *tyrann-iz-ing* raidės *e* priesagoje nėra (Creutz ir kt. 2005: 5).

Antrojoje versijoje *Morfessor 2.0*, be morfologinio žodžių skaidymo, numatytas ir sakinių skaidymas į segmentus. Todėl šioje versijoje įvedami nauji terminai. Mažiausias vienetas, kurio nebegalima toliau skaidyti, yra atomas (pvz., raidė). Atomų seka vadinama dariniu (pvz., žodis). Atomų seka, esanti darinio viduje, žymima kaip konstrukcija (pvz., morfema). Laikoma, kad konstrukcijos darinyje pasitaiko nepriklausomai. Versija *Morfessor 2.0* apima ir darinių ribų tikimybes. Joje naudojamas pusiau valdomas (angl. *semi-supervised*) metodas, kurio darbo eiga parodyta 53 pav. (Smit ir kt. 2014: 24).



53 pav. Morfologinio analizatoriaus *Morfessor 2.0* darbo eiga (parengta pagal Smit ir kt. 2014: 24)

Apmokymo duomenys yra neanotuotas tekstas ir šalia pateikiamas tam tikras kiekis anotuotų žodžių. Analizuodamas abiejų tipų apmokymo duomenis, *Morfessor* susikuria modelį, pagal kurį atlieka jam pateiktas užduotis – apdoroja tekstyną. Iš esmės tai yra tekstyno žodžių segmentavimas. Kad būtų galima įvertinti, kokiu tikslumu jis dirba, jam pateikiami auksinio standarto sakiniai – rankomis žmogaus anotuoti tekstai (Nothman, Murphy, Curran 2009: 612), turintys labai aukšto tikslumo ir patikimumo duomenis, – ir tikrinama, kiek jų analizė sutampa su *Morfessor 2.0* darbu.

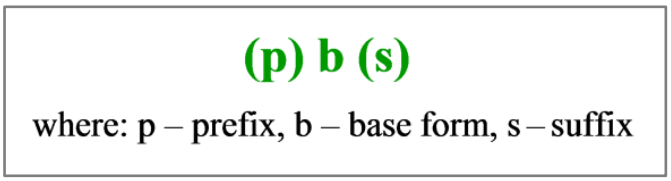
Reikia pasakyti, kad tokiu metodu veikiančių analizatorių darbas daugiausia apsiriboja žodžių suskaidymu į morfemas.

3.2. Žodžio morfeminės struktūros pavaizdavimas

Įvairiose kalbose žodžio struktūra vaizduojama nevienodai. Suomių ir latvių kalbų žodžiai turi po keturias dalis, tačiau jose pateikiama skirtinga informacija. Rusų kalbininkai aprašo žodį kaip turintį dvi dalis. Anglų kalbos žodį sudaro trys dalys. Polisintetinėse kalbose žodis taip pat skaidomas į tris dalis, tačiau jis atitinka prasmę, kuri anglų kalboje paprastai perteikiama keliais žodžiais.

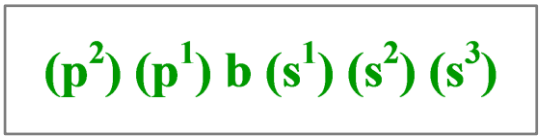
3.2.1. Anglų kalbos žodžio struktūra

Anglų kalbos žodyje išskiriamos trys dalys – priešdėlis, šaknis ir priesaga. 54 pav. pateikta anglų kalbos žodžio formulė (Adedimeji 2005: 10).



54 pav. Anglų kalbos žodžio formulė (parengta pagal Adedimeji 2005: 10)

Šaknis yra būtina, o priešdėlio ir priesagos gali ir nebūti. Tai parodoma įrašant juos skliaustuose. Anglų kalbos žodžių struktūros galimos tokios: *b*, *pb*, *bs*, *pbs*, be to, žodžiai gali turėti daugiau nei vieną priešdėlį ir daugiau nei vieną priesagą. Sudurtiniai žodžiai turi daugiau nei vieną šaknį. Anglų kalbos žodžio formulė gali būti išplėsta, tokia žodžio struktūra pateikta 55 pav.



55 pav. Išplėsta anglų kalbos žodžio struktūra (parengta pagal Adedimeji 2005: 11)

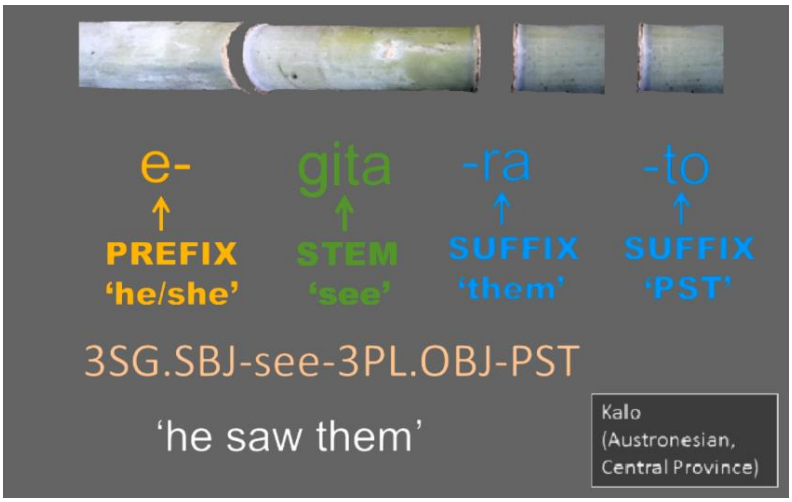
Tačiau autorius (Adedimeji 2005) nenurodė, ar žodžio struktūra su dviem priešdėliais ir trimis priesagomis yra maksimalus išplėtimas anglų kalboje.

3.2.2. Kalo kalbos žodžių struktūra

Naujojoje Gvinėjoje, Ramiojo vandenyno kalbų, menų ir vertimo institute (Pacific Institute of Languages, Arts and Translation), yra sukurta metodika, kaip morfemomis pavaizduoti ne indoeuropiečių kalbų žodžius. Polisintetinių kalbų grupei priklausančia kalo kalba šneka Austronezijos Centrinės provincijos gyventojai. Joje, kaip ir visose šio tipo kalbose, vienu žodžiu pasakoma tai, ką indoeuropiečiai kartais išreiškia net visu sakiniu. Žodis skaidomas į priešdėlį, kamieną (mūsų supratimu, tai labiau atitiktų šaknį) ir priesagų sritį, kuri gali apimti vieną jų ar daugiau. 56 pav. parodytas iš keturių morfemų susidedantis kalo kalbos žodis *egitarato*, kuris atitinka tris anglų kalbos žodžius *he saw them*. 57 pav. pateikta *egitarato* morfeminė analizė

(2 interneto nuoroda[74]): 3SG.SBJ (trečiojo asmens vienaskaitos įvardis – veiksnys), 3PL.OBJ (trečiojo asmens daugiskaitos įvardis – papildinys), PST (būtasis laikas).

56 pav. Kalo kalbos žodžio pavyzdys (2 interneto nuoroda, 4 min. 11 sek.)



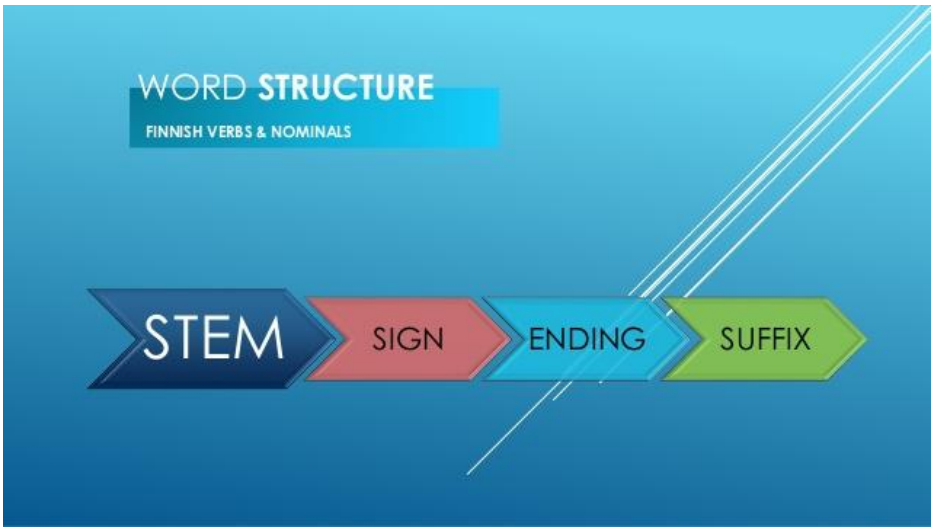
57 pav. Morfemomis išskaidytas kalo kalbos žodis *egitarato* (2 interneto nuoroda, 5 min. 8 sek.)

3.2.3. Suomių kalbos žodžio struktūra

Suomių kalbos žodžio struktūrą sudaro keturios dalys: kamienas (angl. *stem*), požymis (angl. *sign*), galūnė (angl. *ending*) ir priesaga (angl. *suffix*). 58 pav. parodyta suomių kalbos veiksmažodžių ir vardažodžių morfeminė struktūra (46 interneto nuoroda[75]).

[74] Prieiga internete: https://www.youtube.com/watch?v=8ypQq5MvT24. [žiūrėta 2022-11-22].

[75] Prieiga internete: https://www.slideshare.net/riortamm/word-structure-42667712. [žiūrėta 2022-11-22].



58 pav. Suomių kalbos žodžio struktūra (46 interneto nuoroda)

Stem dalis nurodo šaknį ar lemą. *Sign* dalis žymi nuosaką, pvz., tariamąją, arba skaičių. *Ending* dalis pateikia morfologinę kaitybinę informaciją, pvz., asmenį, linksnį. *Suffix* gali turėti savybinį požymį arba reikšti *taip / ne* klausimą ir pan. (46 interneto nuoroda, 2 skaidrė).

3.2.4. Latvių kalbos žodžio struktūra

Latvių kalbos žodžio struktūrą taip pat sudaro keturios dalys, tačiau informacija, nurodoma kiekvienoje dalyje, labai skiriasi nuo suomių kalbos žodžių. 59 pav. pateiktas bendras visų vienašaknių žodžių formatas (Levane, Spektors 2000: 1095).

{<prefix>* <root> <suffix>* [ending]>}

59 pav. Bendras latvių kalbos vienašaknių žodžių formatas
(parengta pagal Levane, Spektors 2000: 1095)

Priešdėlio (angl. *prefix*) ir priesagos (angl. *suffix*) pažymėjimas žvaigždute rodo, kad jie gali pasikartoti kelis kartus arba jų iš viso gali nebūti. Laužtiniuose skliaustuose parašyta galūnė (angl. *ending*) reiškia, kad ji yra nebūtina.

3.2.5. Rusų kalbos žodžio struktūra

Rusų kalbos žodžių struktūroje išskiriamos dviejų tipų morfemos: šaknies (ji žymima raide *R*) ir afiksų, kuriems pažymėti naudojama raidė *x*. Vienašaknių žodžių morfeminės struktūros pavyzdžiai parodyti 60 pav. (47 interneto nuoroda[76]). Vienašakniai žodžiai gali turėti nuo vienos iki aštuonių morfemų.

R *вдруг*;
xR *ни-где*
Rx *вод-а*;
xRx *при-город-ы*;
xRxx *в-пят-ер-ом;*
Rxxx *труд-и-ть-ся*;
Rxxxx *рыб-ач-и-вш-ий*;
xxRxxx *пере-во-оруж-и-вш-ий*;
xxxRxxx *по-на-вы-пис-ыва-л-и*;
xxxRxxxx *по-на-вы-дёрг-ива-л-о-сь.*

60 pav. Vienašaknių rusų kalbos žodžių morfeminės struktūros pavyzdžiai (47 interneto nuoroda)

Daugiašakniuose žodžiuose gali būti ir didesnis morfemų skaičius. 61 pav. pateiktas žodžio, turinčio 11 morfemų, pavyzdys. Dauguma rusų kalbos žodžių turi nuo dviejų iki keturių morfemų (47 interneto nuoroda).

RxRxxxRxxxx
пыл-е-влаг-о-не-про-ниц-а-ем-ость-ю

61 pav. Rusų kalbos daugiašaknio žodžio morfeminės struktūros pavyzdys (47 interneto nuoroda)

[76] Prieiga internete: https://www.langust.ru/rus_gram/rus_gr03.shtml [žiūrėta 2022-11-22].

3.2.6. Lietuvių kalbos žodžio struktūra

Ilgiausias *Dabartinės lietuvių kalbos tekstyne* rastas žodis sudarytas iš devynių morfemų, pvz., *ne-į-si-par-ei-g-o-k-ite* (Rimkutė, Kazlauskienė, Raškinis 2010: 95). Tačiau tai neparodo visų galimų lietuvių kalbos atvejų. Šį žodį, pateiktą kaip pavyzdį, nesunkiai galima papildyti dar vienu priešdėliu ir gauti jau 10 morfemų: *ne-be-į-si-par-ei-g-o-k-ite* [*tiek daug*]. Bandymų grupuoti žodžius į modelius pagal morfeminę sudėtį būta dar praeitame amžiuje. Pagal morfemų skaičių žodyje buvo išskirti septyni tipai – nuo vienmorfemių iki septynių morfemų žodžių (Kuosienė 1986: 100); iš viso sudaryta apie 50 žodžio modelių. Atliekant žodžių morfeminės struktūros tyrimus statistiniais metodais, buvo analizuojamas 310 000 žodžių apimties tekstynas. Skaitvardžiams ir įvardžiams buvo sudaryta 10 modelių, daiktavardžiams ir būdvardžiams – 63, veiksmažodžiams – 116. Nustatyti šeši dažniausiai pasitaikantys morfeminės struktūros modeliai, kurie analizuotame tekstyne sudaro apie 78 proc. visų kaitomų kalbos dalių. Jie parodyti 62 pav. Trys pirmieji sudaro atitinkamai 41 proc., 17 proc. ir 8 proc. visų atvejų (Rimkutė, Kazlauskienė, Utka 2016: 177). Taigi, kaip matyti iš paveikslėlio, beveik pusė tekstyne esančių žodžių sudaryti iš dviejų morfemų. Autoriai pateikia išvadą, kad dauguma lietuvių kalbos žodžių sydaryti iš 2-4 morfemų. Panašūs duomeys nurodyti ir apie rusų kalbos žodžius.

šaknis + galūnė	41 %
šaknis + darybinė priesaga + galūnė	17 %
priešdėlis + šaknis + galūnė	8 %
šaknis + darybinė priesaga + kaitybinė priesaga + galūnė	6 %
šaknis + kaitybinė priesaga + galūnė	3 %
priešdėlis + šaknis + darybinė priesaga + galūnė	3 %

62 pav. Šeši būdingiausi lietuvių kalbos veiksmažodžio modeliai (parengta pagal Rimkutė, Kazlauskienė, Utka 2016: 177)

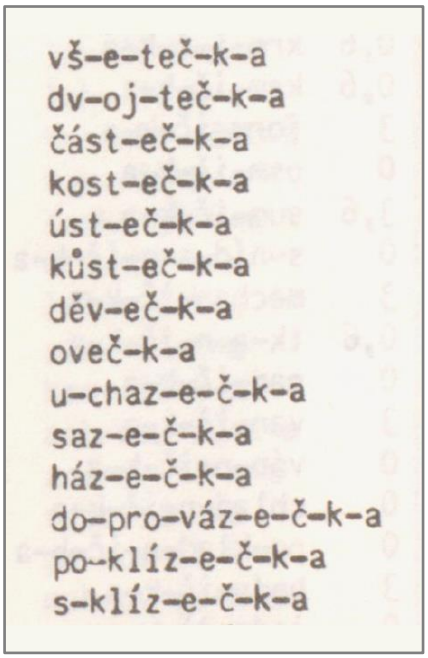
Apibendrintas lietuvių kalbos žodžio formatas pateikiamas 6.1.1.2 poskyryje.

3.3. Morfemikos kompiuterizavimo darbai

Lietuvių kalbos kompiuterizavimo darbai morfemikos srityje pradėti labai neseniai – pirmasis viešai internete prieinamas morfemikos žodynas pasirodė tik 2011 m. Todėl trumpai bus apžvelgtos ir kitų šalių publikacijos, kuriose aprašomas žodžių skaidymas į morfemas.

3.3.1. Kitų kalbų morfemikos darbai

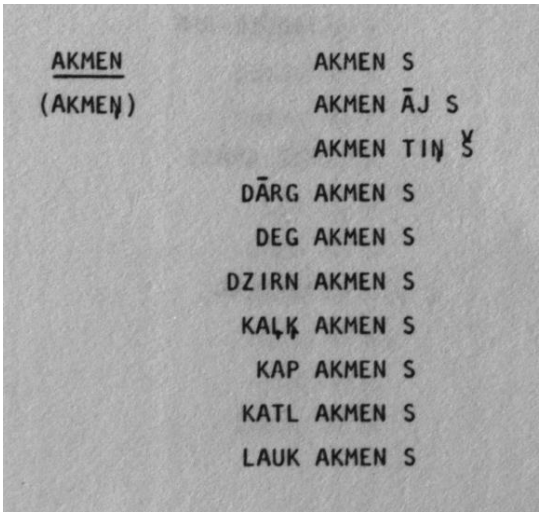
Vienas pirmųjų morfemikos žodynų buvo išleistas Čekijoje 1975 m. Tai atgalinis žodynas, kurio apimtis – apie 64 000 morfemomis išskaidytų žodžių (Slavičkova 2018: 41). Žodžiai išrūšiuoti pagal abėcėlę nuo žodžio pabaigos, bet išdėstyti į stulpelį lygiuojant pagal pirmą morfemą, t. y. žodžio pradžią. Šio žodyno pavyzdys (Slavičkova 2018: 117) pateiktas 63 pav. Toks žodžių išdėstymo būdas pasirinktas ir sudarant visus lietuvių kalbos morfemikos žodynus (Rimkutė, Kazlauskienė, Raškinis 2011).



63 pav. *Čekų atgalinio žodyno* fragmentas (parengta pagal Slavičkova 2018: 117)

Tai pats neinformatyviausias morfemų pavaizdavimo būdas, nes nėra informacijos apie tai, kuo žodis prasideda: priešdėliu ar šaknimi ir, jei priešdėliu, tai neaišku, kiek jų yra.

Latvių kalbos žodžių darybos žodynas išleistas 1985 m. (Metuzale-Kangere 1985). Jame morfemos atskiriamos viena nuo kitos tarpais ir šaknis išdėstoma stulpeliu. Šio žodyno pavyzdys (Metuzale-Kangere 1985: 4) pateiktas 64 pav. Toks pavaizdavimo būdas yra informatyvesnis, nes tiksliai žinoma šaknis, tačiau turėtų kilti problemų, kai žodžiai yra sudurtiniai, nes jų šaknys yra dvi ar net trys. Tada nebelieka priemonių, kaip jas atskirti nuo priesagų ar galūnės (pvz.: *kaipmat*, *tąsyk* ir kt. – antra šaknis užimtų galūnės poziciją).



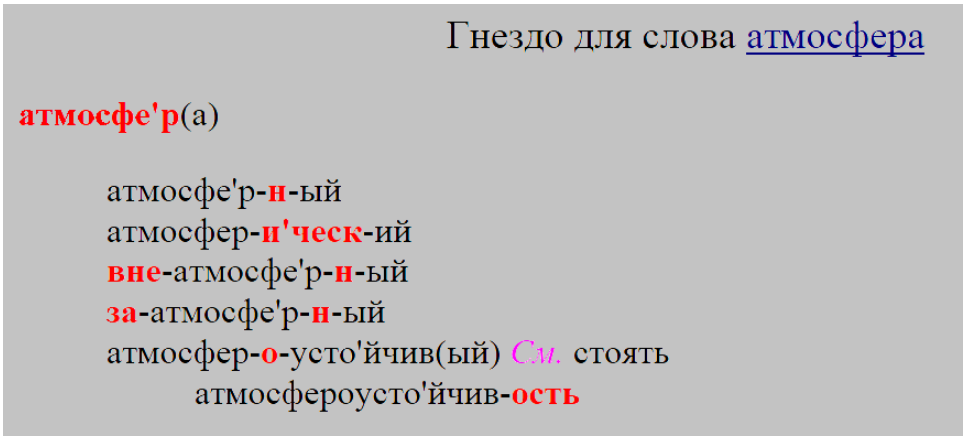
64 pav. *Latvių kalbos darybinio žodyno* pavyzdys: morfeminis žodžių išskaidymas (parengta pagal Metuzale-Kangere 1985: 4)

Vėliau išleistame *Čekų kalbos žodžių darybos žodyne* šaknis išskiriama pasviraisiais brūkšneliais. 65 pav. pateiktas šio žodyno pavyzdys (Sedlaček 2004: 1280). Visos morfemos taip pat atskiriamos tarpais.

```
          /děd/
          /děd/ a
          /děd/ eč ek
          /děd/ ek
           ...
     pra /děd/
     pra /děd/ eč ek
pra pra /děd/
pra pra /děd/ eč ek
           ...
```

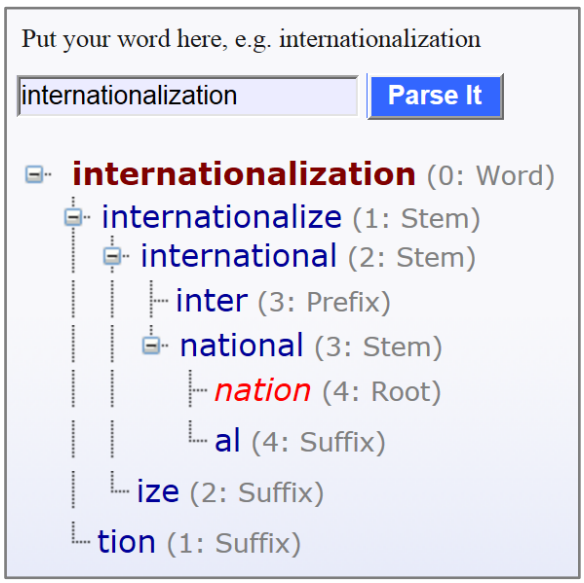
65 pav. *Čekų kalbos darybinio žodyno* pavyzdys: morfeminis žodžių išskaidymas (parengta pagal Sedlaček 2004: 1280)

Rusų kalbos morfeminio žodžių skaidymo informacija pateikiama greta esančius afiksus vaizduojant skirtingomis spalvomis (48 interneto nuoroda[77]). 66 pav. parodytas žodžio *атмосфера* lizdas.



66 pav. *Rusų kalbos morfeminio žodyno* pavyzdys (48 interneto nuoroda)

Internete buvo viešai prieinamas anglų kalbos analizatorius (49 interneto nuoroda[78]), kuris, nors ir vadinamas morfologiniu, pateikdavo morfeminę žodžio informaciją. 67 pav. parodytas anglų kalbos žodžio *internationalization* analizės rezultatas.



67 pav. Anglų kalbos žodžio *internationalization* morfeminė analizė (49 interneto nuoroda)

[77] Prieiga internete: http://old.kpfu.ru/infres/slovar1/begall.htm [žiūrėta 2021-12-02].
[78] Prieiga internete: http://nlpdotnet.com/services/Morphparser.aspx [žiūrėta 2016-01-22].

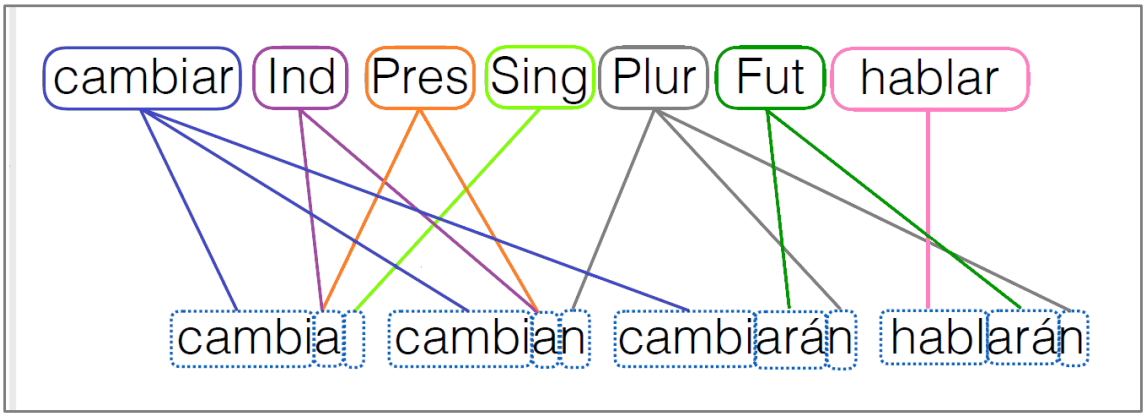
3.3.2. Automatinė morfeminė analizė

Morfologiškai anotuojant tekstą paprastai žodžiams nurodomi tokie požymiai kaip linksnis, giminė, skaičius, laikas ir kt. Tačiau šie požymiai nesusieti su konkrečia raidžių grupe žodyje. Pavyzdžiui, ispanų kalbos žodžio *asignados* duomenys pateikti 68 pav. (Sifverberg, Hulden 2017: 141).

```
Spanish  asignados  Verb|Lemma=asignar|Gender=Masc|Number=Plur|Tense=Past|VerbForm=Part
```

68 pav. Žodžio *asignados* morfologiniai duomenys (parengta pagal Silfverberg, Hulden 2017: 141)

Kolorado universiteto (University of Colorado) mokslininkai pasiūlė metodą, kaip anotuotame tekstyne susieti morfologinę informaciją su atitinkama žodžio dalimi: *asign-* atitinka kamieną, *-ad-* rodo, kad tai yra dalyvis, *-o-* žymi vyriškąją giminę ir *-s* – daugiskaitą. Metodas buvo taikomas trims kalboms: švedų, suomių ir ispanų. Ranka sužymėti duomenys naudoti pusiau valdomam mokymuisi. Kadangi dauguma morfologinių požymių apmokymo duomenyse pasitaikė labai daug kartų, buvo galima sukurti modelį, kuris nustatytų sistemingą atskirų morfemų atitiktį tam tikrai žodžio daliai. Žinoma, reikia įvertinti alomorfų atvejus, pvz., anglų kalboje daugiskaitą gali žymėti trys alomorfai: *-s* (*cat* – *cats*), *-es* (*fox* – *foxes*) ir nulinis atvejis (*sheep* – *sheep*). Alomorfas *-s* anglų kalboje atlieka dvi funkcijas: žymi daugiskaitą ir esamojo laiko trečiąjį asmenį (*dog* – *dogs* ir *take* – *takes*). 69 pav. pateiktas morfemų ir morfologinių požymių atitikmenų pavyzdys (Silfverberg, Hulden 2017: 141).



69 pav. Morfologinių požymių priskyrimas morfemoms
(parengta pagal Silfverberg, Hulden 2017: 141)

Kaip matyti iš paveikslėlio, kartais tam tikro požymio atitikmuo gali būti ir nulinė morfema, šiuo atveju – vienaskaitos. Pasiektas tikslumas: suomių kalbai – 90,62, švedų kalbai – 97,66, ispanų kalbai – 79,54.

3.3.3. Morfemikos kompiuterizavimo darbai, atlikti Lietuvoje

Pirmasis stambus lietuvių kalbos morfemikos kompiuterizavimo darbas buvo atliktas Matematikos ir informatikos institute 1992 m. Sukurtoje *Žodžių darybos ir morfemų duomenų bazėje* (ŽDMDB) sukaupta gana išsami informacija apie morfemas: kiekvienos rūšies morfemos užrašomos skirtingu šriftu (Murmulaitytė 2012: 96–97). Pavyzdžiui, žodis *tikimybinis* vaizduojamas taip, kaip parodyta 70 pav.: šaknis – pastorintu šriftu, galūnė – paprastuoju, priesagos – pasvirusiuoju šriftu (Murmulaitytė 2012: 98). Jei yra kelios priesagos, jos atskiriamos tarpeliais. Darybinei priesagai žymėti naudojamos didžiosios raidės. Šalia pateikiamas ir pamatinis žodis. 71 pav. parodytas žodžio su priešdėliu – *užjūrinis* – pavyzdys. Priešdėlis užrašomas pasvirusiuoju pabrauktu šriftu.

bdv.	**tik** *im yb IN* is	tikimybė dkt.

70 pav. Žodžio *tikimybinis* pavaizdavimas *Žodžių darybos ir morfemų duomenų bazėje* (Murmulaitytė 2012: 98)

bdv.	*už* **jūr** *IN* is	užjūris dkt.

71 pav. Žodžio *užjūrinis* pavaizdavimas *Žodžių darybos ir morfemų duomenų bazėje* (Murmulaitytė 2012: 98)

Šioje duomenų bazėje sukaupta daug tikrai vertingos informacijos. Tik labai blogai, kad ji nėra viešai prieinama – ja tegali naudotis tik patys autoriai. Ir panašu, kad darbai nėra tęsiami.

2011 m. pasirodė viešai internete prieinamas morfemikos žodynas, kuris buvo sukurtas VDU Kompiuterinės lingvistikos centre remiantis tekstynu ir teapima tik jame esančius žodžius. Morfemos viena nuo kitos atskiriamos brūkšneliais, tačiau nepateikiama jokios informacijos apie morfemos tipą, todėl kartais vienodai

pavaizduojami skirtingos morfeminės struktūros žodžiai. Vėliau šio žodyno pagrindu buvo sukurta ir morfemikos duomenų bazė.

3.3.3.1. Morfemikos žodynas

VDU darbai, atlikti tyrinėjant morfeminę struktūrą, apima 310 000 žodžių (apie 70 000 skirtingų žodžių formų) analizę (Rimkutė, Kazlauskienė, Raškinis 2011a: 7). Imti kelių skirtingų sričių tekstai: moksliniai, grožinės literatūros, publicistikos, administraciniai. Morfemų ribos sužymėtos ranka. Rezultatai pateikiami trijų tomų žodyne, kuris apima bendrinius daiktavardžius, būdvardžius, įvardžius, skaitvardžius ir veiksmažodžius (Rimkutė 2017: 38). Prie kiekvieno žodžio, išskaidyto morfemomis, nurodoma jo pradinė forma (lema), morfologiniai duomenys ir dažnumas tekstyne.

Labai didelis privalumas – kad išleistos trys šio žodyno versijos: abėcėlinis (Rimkutė, Kazlauskienė, Raškinis 2011a, 2011b, 2011c), dažninis (Rimkutė, Kazlauskienė, Raškinis 2011d, 2011e, 2011f) ir atbulinis (Rimkutė, Kazlauskienė, Raškinis 2011g, 2011h, 2011j). Todėl sudarytos sąlygos tyrinėti lietuvių kalbos žodžių morfeminę struktūrą įvairiais aspektais.

Abėcėlės tvarka pateiktame žodyne galima gana greitai rasti žodį ir sužinoti jo vartojimo dažnį. Tokiame žodyne labai gerai matyti giminiškų žodžių morfeminė struktūra, pvz.: *adresas*, *adresatas*, *adresavimo* ir pan. Taip pat galima analizuoti pasirinkto žodžio formų vartojimo dažnumą, pvz.: *gamyba* (41), *gamybą* (32), *gamybai* (24), *gamyboje* (16), *gamybos* (263).

Atgaliniame žodyne žodžiai surikiuoti pagal abėcėlę skaitant žodį nuo pabaigos. Jame lengvai pastebimi darybiniai tipai: greta pateikti to paties afikso žodžiai, pvz.: *stebėtojai*, *nugalėtojai*, *laimėtojai*, *tyrinėtojai*, *prižiūrėtojai*; arba *lizdavietė*, *slaptavietė*, *statybvietė*, *prekybvietė* ir kt.

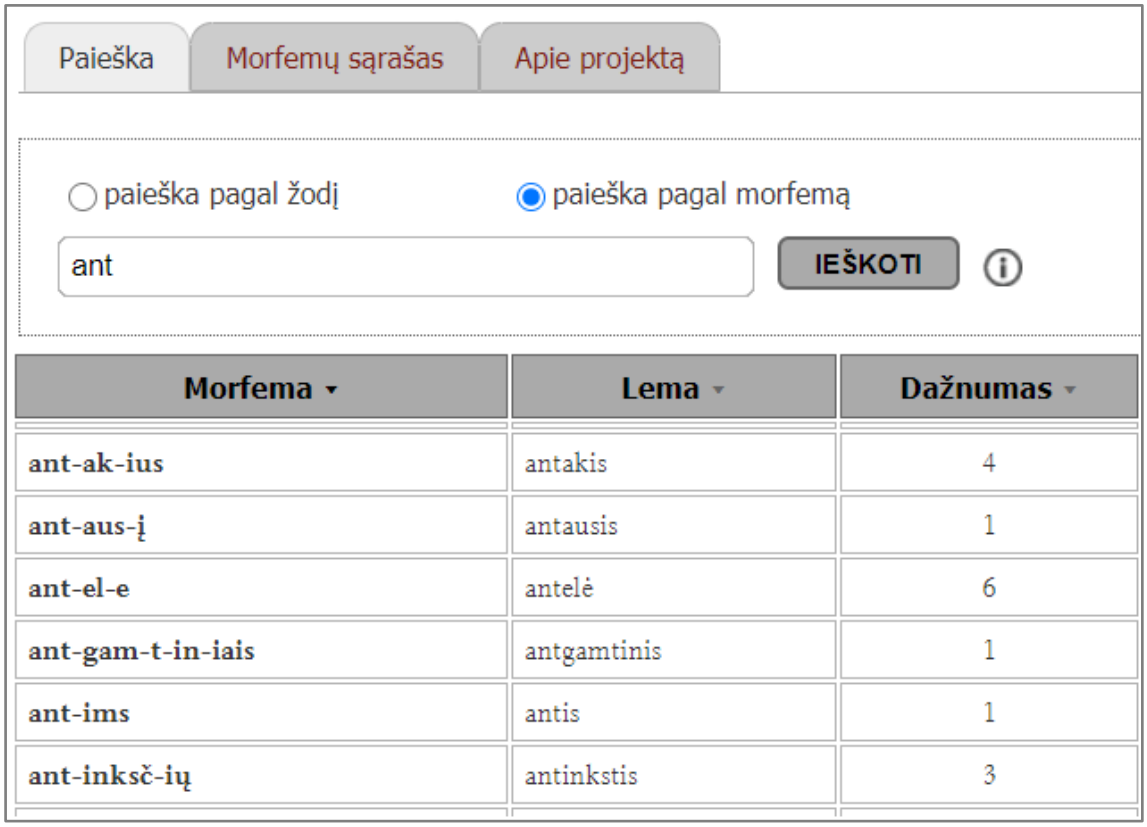
Mažėjančio dažnio tvarka pateiktame žodyne atsispindi tirtų žodžių statistiniai duomenys: dažniausiai ir rečiausiai pasitaikantys žodžiai. Galima gretinti žodžius pagal vartojimo polinkius, pvz.: *darbo* (1 136), *darbuotojų* (579), *metų* (528), *kalbos* (477), *žmogus* (372) ir t. t. (Rimkutė 2017: 39–41).

Kaip vieną iš trūkumų galima būtų paminėti informacijos apie morfemos tipą stoką. Nors žodyno aprašyme sakoma, kad *-un-* laikoma priesaga žodyje *šunį* (Rimkutė, Kazlauskienė, Raškinis 2011: 7), tačiau žodyne jis pateikiamas tokios pat struktūros, kaip ir žodis *sutemos,* t. y. trys raidžių rinkiniai, atskirti brūkšneliais: *š-un-s* (Rimkutė, Kazlauskienė, Raškinis 2011a: 686) ir *su-tem-os* (Rimkutė, Kazlauskienė, Raškinis

2011a: 665). Abu žodžiai sudaryti iš trijų morfemų, tačiau visai nėra informacijos apie tai, kad žodyje *šuns* pirma morfema yra šaknis, antra – priesaga, o žodyje *sutemos* pirma morfema yra priešdėlis, o antra – šaknis. Žodžiai išdėstyti žodyne lygiuojant juos pagal pirmą raidę, t. y. žodžio pradžią, todėl nėra jokio vizualaus skirtumo tarp žodžių, prasidedančių priešdėliu, ir žodžių, prasidedančių šaknimi.

3.3.3.2. *Lietuvių kalbos morfemikos duomenų bazė*

2013 m. morfemikos žodyno pagrindu buvo sukurta viešai prieinama internete *Lietuvių kalbos morfemikos duomenų bazė* (50 interneto nuoroda[79]). Žodis joje skaidomas į morfemas tuo pačiu principu – atskiriant jas brūkšneliais, kaip tai buvo daroma žodyne. Todėl žodžiai *antakius* ir *antele* pavaizduoti vienodai – kaip turintys tris morfemas ir pirmoji jų yra *ant* (72 pav.). Nėra jokios informacijos apie tai, kad žodyje *antakius* ši morfema (*ant*) yra priešdėlis, o žodyje *antele* ta pati morfema (*ant*) yra šaknis. Visos morfemos *ant* išdėstytos viename stulpelyje.



Morfema	Lema	Dažnumas
ant-ak-ius	antakis	4
ant-aus-į	antausis	1
ant-el-e	antelė	6
ant-gam-t-in-iais	antgamtinis	1
ant-ims	antis	1
ant-inksč-ių	antinkstis	3

72 pav. Žodžių *antakius* ir *antele* pavaizdavimas *Lietuvių kalbos morfemikos duomenų bazėje* (50 interneto nuoroda)

[79] Prieiga internete: https://klc.vdu.lt/morfema/ [žiūrėta 2022-11-22].

Lygiai taip pat vienodai vaizduojami žodžiai, turintys dvi šaknis bei galūnę, ir žodžiai, turintys šaknį, priesagą ir galūnę, pvz., *laik-rod-is* ir *laik-men-oje* (Rimkutė, Kazlauskienė, Raškinis 2011: 331–332), abu jie sudaryti iš trijų morfemų ir antra šaknis *rod-* vizualiai niekuo nesiskiria nuo priesagos *-men-*.

Bene didžiausias šios duomenų bazės trūkumas yra tas, kad negalima paieška pagal morfemos tipą. Patys autoriai duomenų bazės aprašyme nurodo: „Internete prieinamoje duomenų bazėje ribos tarp morfemų žymimos brūkšneliais. Dėl šios priežasties kol kas negalima ieškoti tam tikrų rūšių morfemų, pvz., šaknų, priešdėlių, galūnių ir pan.“ (50 interneto nuoroda[80], skirtukas *Paieška -> Plačiau*).

3.4. Skyriaus išvados

Atliekant automatinę morfologinę analizę, naudojami du pagrindiniai metodai: taisyklėmis pagrįstas ir statistinis. Taisyklėmis pagrįstas metodas reikalauja daug labai aukštos kvalifikacijos specialistų darbo, bet jam nereikia jokių papildomų išteklių, tokių kaip ranka anotuoti tekstynai. Šiuo metodu veikianti programinė įranga pateikia labai tikslią analizę, bet lieka daugiareikšmiškumas.

Statistiniais metodais dirbančioms sistemoms reikia palyginti nedidelio kiekio žmogaus anotuotų sakinių ir jos gali anotuoti labai didelės apimties tekstynus. Tačiau rezultatai gaunami su tam tikra tikimybe, t. y. absoliutaus tikslumo čia nepasiekiama.

Statistiniai metodai gana gerai tinka, kai norima žodžius suskaidyti morfemomis. Ypač jie vertingi tais atvejais, kai reikia sudaryti mažai tyrinėtų kalbų gramatikas arba kai nebelikę jomis kalbančių žmonių.

VDU parengtas lietuvių kalbos morfologinis analizatorius, veikiantis taisyklėmis pagrįstu metodu, buvo kuriams *Hunspell* platformoje. Statistiniais metodais lietuvių kalbos morfologinę analizę atlieka *UDPipe* lietuvių kalbos modulis.

Atskiros kalbos labai nevienodai skaido žodžius morfemomis. Dažniausiai išskiriamos keturios (suomių, latvių kalbose) ar trys (anglų kalboje) žodžio dalys, kartais – tik dvi (rusų kalboje). Tačiau, net ir skaidant žodį į tą patį skaičių dalių, kiekvienoje dalyje pateikiama informacija įvairiose kalbose dažniausiai būna ne tokia pat. Polisintetinėms kalboms, pvz., kalo kalbai, taip pat siūlomas morfeminis

[80] Prieiga internete: https://klc.vdu.lt/morfema/ [žiūrėta 2022-11-22].

skaidymas, nors šių kalbų žodis dažniausiai atitinka pasakymus, kurie anglų kalboje išreiškiami keliais žodžiais ar net visu sakiniu.

Kitų kalbų (čekų, latvių) morfemikos žodynai pasirodė anksčiau nei lietuvių kalbos morfemikos žodynas. Jis buvo parengtas tekstyno pagrindu ir yra išleistos trys jo versijos elektronine forma: abėcėlinis, dažninis ir atbulinis.

Pirmoji morfemikos duomenų bazė, sukurta MII, apima išsamią informaciją apie morfemas, įskaitant ir morfemos tipą, tačiau jos duomenys nėra laisvai prieinami. Internete publikuojama VDU *Lietuvių kalbos morfemikos duomenų bazė* pateikia žodžius, išskaidytus morfemomis atskiriant jas viena nuo kitos brūkšneliu, tačiau informacijos apie morfemos tipą ši duomenų bazė neturi.

4. SINTAKSĖS KOMPIUTERIZAVIMAS

„Mes galime sudaryti bei pasakyti sakinius, kurių niekada anksčiau nesame ištarę ar girdėję, ir tuos sakinius supranta kiti žmonės, kuriems jie yra iš viso nauji, nežinomi, niekada anksčiau negirdėti. Tačiau mūsų laisvė sakyti nauja yra įsprausta į tam tikrus rėmus. Visos kalbos turi sintaksę, kuri uždeda griežtas toje kalboje vartojamų modelių ribas“ (Winograd 1983: 35).

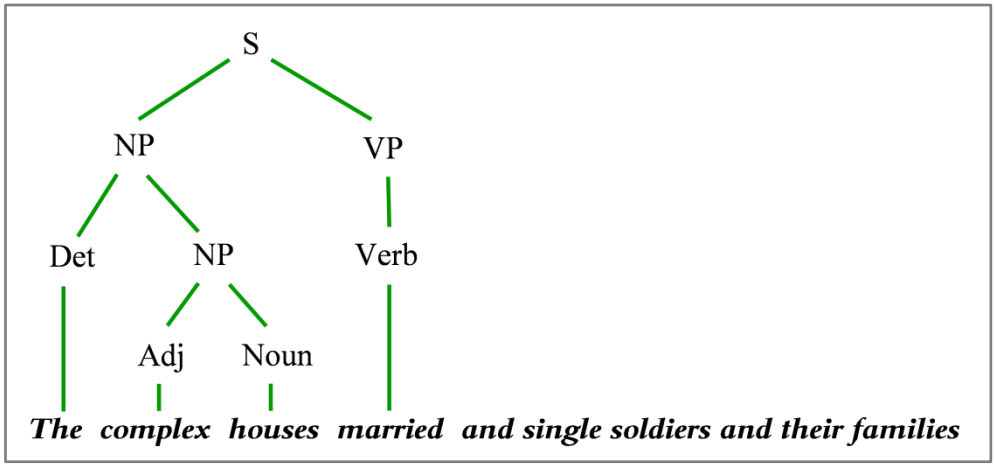
Žodis *sintaksė* kilęs iš graikų kalbos *syntaxis*, kurio pirminė reikšmė buvo sąryšis, išsidėstymas, išsirikiavimas, sąveika. Graikai jį vartojo karo moksluose, kalbėdami apie kariuomenę; vėliau jis pateko į logiką, o po to – į gramatiką (Labutis 2002: 8). Sakinio sintaksinė struktūra rodo, kaip žodžiai sakinyje yra susiję vienas su kitu (Allen 1987: 9).

Atrodytų, kad sintaksė yra gana formalus dalykas ir kompiuterizuoti ją neturėtų būti labai sunku. Tačiau tik nedaugeliui pasaulio kalbų pavyko pasiekti neblogų rezultatų, nors šioje srityje dirba kelių sričių specialistai. „Kompiuterinės lingvistikos tyrinėtojai domisi, kaip žmonės vartoja bei suvokia kalbą. Jų tikslas – sukurti kalbos apdorojimo mechanizmo, esančio žmoguje, kompiuterinį analogą“ (Allen 1987: 2).

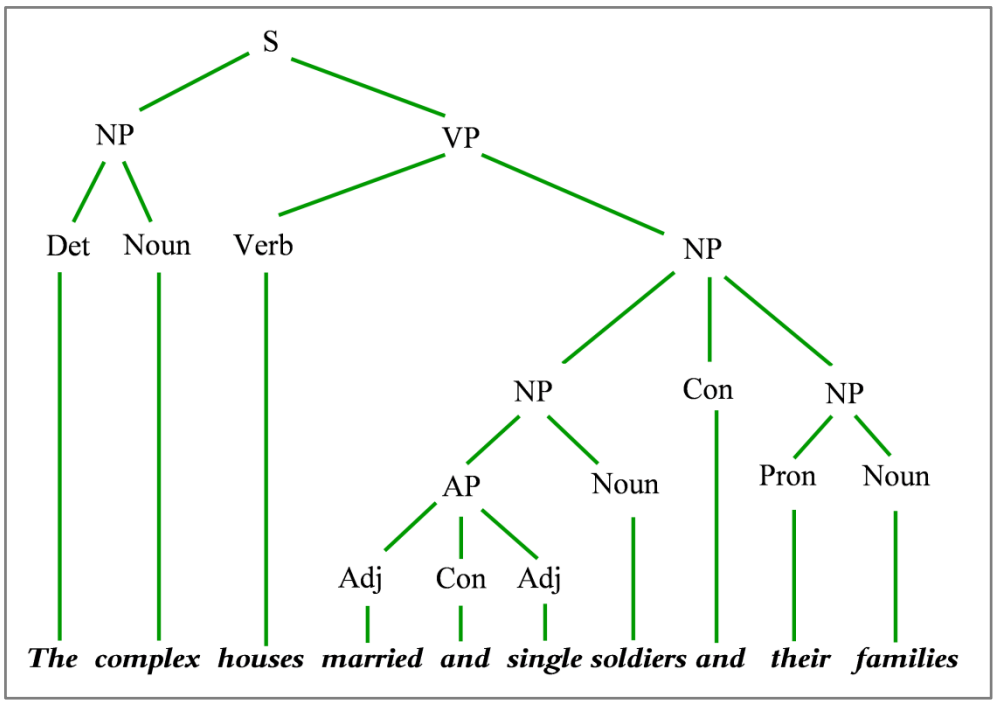
Psicholingvistinių eksperimentų duomenys kartais naudojami patvirtinant ar atmetant įvairias hipotezes apie kalbą, kurias iškelia teorinės ar kompiuterinės lingvistikos tyrinėtojai (Allen 1987: 2). Tokios hipotezės pavyzdys galėtų būti Rusijos mokslininkų iškelta mintis, kad, remiantis tekste pavartotų sakinių sintaksinėmis struktūromis, galima nustatyti autorių pagal jo individualų sintaksinį braižą (Белецкая 1983: 37).

Įdomi kalbos suvokimo ypatybė yra ta, kad painūs, sudėtingi pasakymai kartais gali būti labai sunkiai suprantami, o, braižant pieštuku ant popieriaus sakinio struktūrą, galima daug lengviau atsekti, kas su kuo yra susiję (Winograd 1983: 145), pvz., sakinys *The complex houses married and single soldiers and their families* iš pirmo žvilgsnio gali atrodyti nesąmoningas, jei tariniu laikysime žodį *married* (51 interneto nuoroda[81]). 73 pav. parodyta tokios sakinio analizės pradžia. Tačiau jei tarinio funkciją priskirsime žodžiui *houses*, sakinys taps aiškus. 74 pav. pateikta viso sakinio analizė.

[81] Prieiga internete: https://en.wikipedia.org/wiki/Garden-path_sentence [žiūrėta 2022-11-22].



73 pav. Sakinio analizė veiksmažodžiu laikant *married* (51 interneto nuoroda)



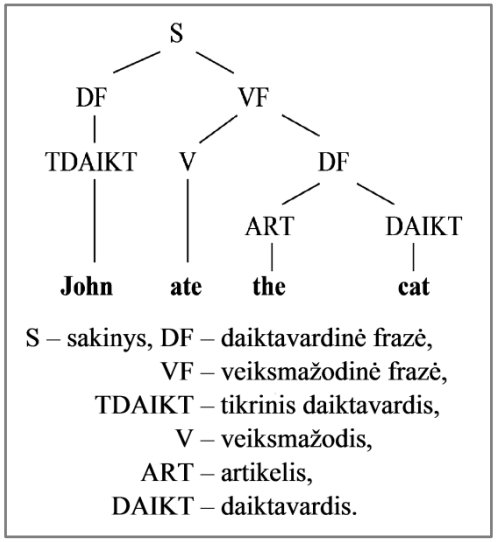
74 pav. Sakinio analizė veiksmažodžiu laikant *houses* (51 interneto nuoroda)

4.1. Sakinio sintaksinės struktūros pavaizdavimas

Sakinio sintaksinei struktūrai pavaizduoti iš pradžių buvo sukurti du pagrindiniai iš principo skirtingi metodai – frazių (Chomsky 1956) ir priklausomybių (Tesnière 1959). Abu jie yra sąlygoti formalių gramatikų reikalavimo turėti vieną pradinį simbolį (plačiau apie tai žr. 4.2.1.1 poskyryje). Pirmoji buvo sukurta frazių gramatika, pradiniu simboliu pasirinkusi sakinį. Vėliau pasirodė priklausomybių

gramatika, kuri pradiniu simboliu laiko tarinį. Pastaruoju metu abu metodai sujungti ir sukurtos universaliosios priklausomybės (Nivre 2014).

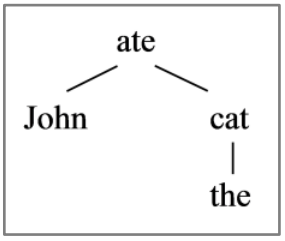
Frazių gramatikoje pradinis simbolis *S* – sakinys – iš pradžių skaidomas į dvi grupes – veiksnio ir tarinio, kaip ir „tradicinėje gramatikoje, kurioje sakinio analizė remiasi subjekto ir predikato dichotomija“ (Holvoet, Mikulskas 2009: 10). Frazių gramatika buvo sukurta anglų kalbai, todėl stengtasi maksimaliai atspindėti šios kalbos ypatybes. Atliekant sintaksinę analizę, čia remiamasi specifiniu anglų kalbos bruožu – griežta žodžių tvarka. Būtent ji buvo pagrindinis kriterijus, naudojamas nustatant sakinio dalis. Taigi, ir pati sintaksinė struktūra yra susijusi su žodžių išsidėstymu sakinyje. Kaip pavyzdį galima pateikti labai paprasto anglų kalbos sakinio *John ate the cat* medį (75 pav.). Pagal anglų kalbos gramatiką veiksnys turi būti pirmoje vietoje, o tarinys – antroje. Todėl sakinio struktūroje pirmiau eina daiktavardinė frazė (DF), o paskui – veiksmažodinė (VF).



75 pav. Anglų kalbos sakinys frazių gramatikoje (parengta pagal Allen 1987: 42)

Tačiau toks pavaizdavimo būdas visai netiko laisvą žodžių tvarką turinčioms kalboms. Todėl Europoje buvo sukurta kita gramatika – priklausomybių, kuri labiau atitiko kalbų, turinčių laisvą žodžių tvarką sakinyje, specifiką: „[...] priklausomybių gramatika pateikia tyrėjui universalesnį aprašymo įrankį“ (Holvoet 2009: 31). Naudojantis ja, žinoma, galima pavaizduoti ir anglų kalbos sakinius, tačiau priklausomybių gramatikoje sakinio struktūra jau nebesusijusi su linijiniu žodžių išsidėstymu. Frazių gramatikoje teigiama, kad sakinys yra sudarytas iš frazių ir ši

struktūra yra svarbi perduodant reikšmę[82] (Winograd 1983: 73). Priklausomybių gramatikoje manoma, kad „[...] sintaksė ne tiek turi grupuoti žodžius į frazes, kiek nustatyti tiesioginius ryšius tarp pačių žodžių“[83] (Kay, Gawron, Norvig 1994: 53). Sakinys *John ate the cat*, kurio struktūra frazių gramatikos metodu pateikta 75 pav., priklausomybių gramatikoje atrodytų taip (76 pav.):



76 pav. Sakinio *John ate the cat* sintaksinė struktūra priklausomybių gramatikoje

Laisvą žodžių tvarką turinčios lietuvių kalbos sakinių sintaksinei struktūrai pavaizduoti labiau tinka būtent šios, priklausomybių gramatikos, metodas. Vaizduojant lietuvių kalbos sakinius frazių gramatikos metodu, kai kuriais atvejais sintaksinė struktūra gali būti labai paini ir sudėtinga (Šveikauskienė 2013: 7–9).

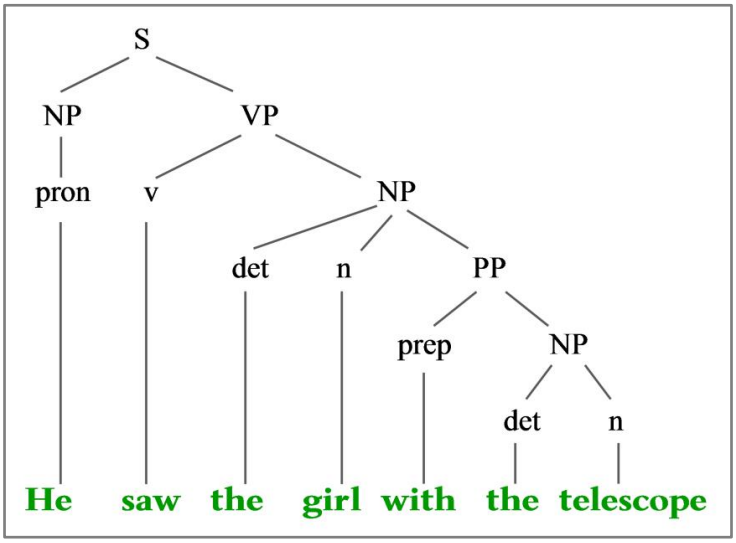
Abiejų gramatikų – frazių ir priklausomybių – svarbiausias skirtumas yra ne tas, kad vienoje jų naudojama dichotominė, kitoje – verbocentrinė sakinio struktūros forma: jos abi turi vieną pradinį simbolį. Pagrindinis skirtumas pastebimas, vaizduojant sakinio struktūrą: frazių gramatikoje sakinys pateikiamas toks, koks jis yra tekste (75 pav.), ir virš sakinio žodžių nurodomos sintaksinės kategorijos (sakinys, daiktavardinė frazė, veiksmažodinė frazė ir kt.). Priklausomybių gramatikoje sintaksinė struktūra nesusijusi su žodžių išsidėstymu sakinyje – mes nematome sakinio tokio, koks jis būna parašytas tekste, ir, jei tai yra mums nežinomos kalbos sakinys, net negalėtume atkurti jo linijinės struktūros, t. y. tokio žodžių išsidėstymo, koks būtinas pagal tos kalbos gramatiką (76 pav.), „todėl norint griežtai aprašyti žodžių tvarką, į sakinio hierarchinės struktūros aprašą reikėtų įtraukti atskirą – linearizacijos – komponentą“ (Holvoet 2009: 24).

Frazių gramatikoje sakinys iš pradžių nebūtinai turi būti skaidomas į dvi dalis, kartais jis iš karto skaidomas į tris, jei kokia nors (pvz., prielinksninė) konstrukcija

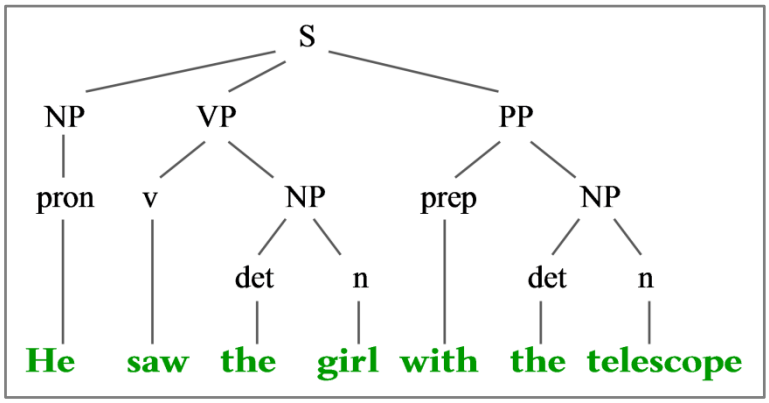
[82] “Sentences are made up of “phrases”, and this structuring is important for how they communicate meaning” (Winograd 1983: 73).

[83] “The job of syntax is not so much to group words into phrases, as to establish direct relationships among the words themselves” (Kay, Gawron, Norvig 1994: 53).

priklauso visam sakiniui, pvz., S -> NP VP PP (Hutchins, Somers 1992: 58). Kartais net apie sakinio prasmę sprendžiama pagal schemą, pvz., sakinio *He saw the girl with the telescope* (Hutchins, Somers 1992: 61) sintaksinį daugiareikšmiškumą panaikina schema – kaip ji bus pateikta, tokia bus ir sakinio prasmė. 77 pav. schemoje parodomas sakinys, kuris reiškia, kad jis matė mergaitę, laikančią savo rankose žiūroną. 78 pav. parodyto sakinio reikšmė: jis su žiūronais stebėjo mergaitę, t. y. matė ją, vadinasi, žiūronas šį kartą buvo jau ne mergaitės, o jo rankose.



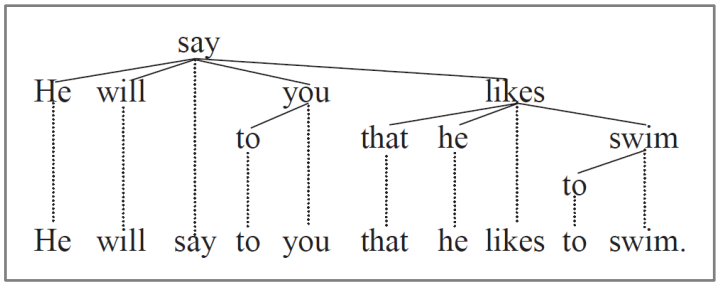
77 pav. Sakinio *He saw the girl with the telescope* sintaksinė struktūra, rodanti, kad žiūronas buvo mergaitės rankose (parengta pagal Hutchins, Sommers 1992: 61)



78 pav. Sakinio *He saw the girl with the telescope* sintaksinė struktūra, rodanti, kad žiūronas buvo jo rankose (parengta pagal Hutchins, Sommers 1992: 61)

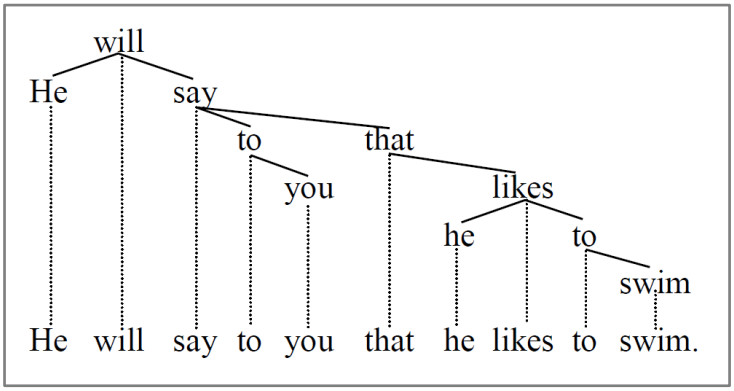
Toks pavyzdys pateiktas literatūroje, nors šią sakinio prasmę turbūt galėtų turėti ir schema, kurioje prielinksninė konstrukcija PP (*prepositional phrase*) priklausytų veiksmažodinei frazei VP (*verb phrase*).

2013 m. internete viešai buvo pateikti šešių kalbų (anglų, vokiečių, švedų, ispanų, prancūzų ir korėjiečių) tekstynai, anotuoti projekto *Universal dependency Treebank* (UDT) metu (McDonald ir kt. 2013: 93). Šiame projekte buvo siekiama tarptautiniu mastu sudaryti daugelio kalbų anotuotus tekstynus, naudojančius tą patį žymėjimą[84] (52 interneto nuoroda[85]), todėl buvo sukurtos universaliosios priklausomybės (angl. *universal dependencies*), kai virš sakinio žodžių tekste parodomi jų ryšiai. 79 pav. pateiktas sakinio *He will say to you that he likes to swim* pavyzdys, kuriame pagrindiniu žodžių junginio dėmeniu laikomi savarankiškos reikšmės žodžiai.



79 pav. Sakinio *He will say to you that he likes to swim* sintaksinė struktūra, pagrindiniu žodžių junginio dėmeniu laikant savarankiškus žodžius (Osborne, Gerdes 2019: 2)

Šiuo metu pasiūlyta pagrindiniu žodžių junginio dėmeniu laikyti pagalbinius žodžius. 80 pav. parodytas tas pats sakinys, tik šį kartą traktuojama, kad pagrindinis dėmuo – pagalbiniai žodžiai (Osborne, Gerdes 2019: 2).

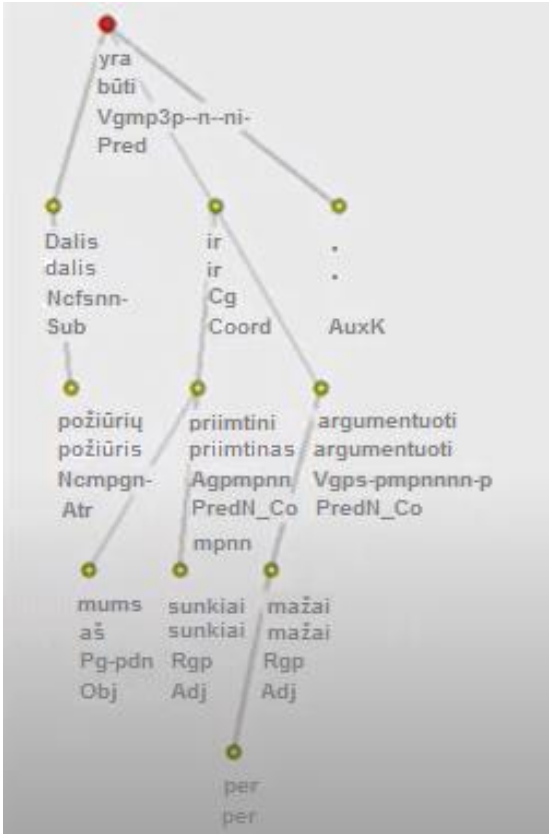


80 pav. Sakinio *He will say to you that he likes to swim* sintaksinė struktūra, pagrindiniu žodžių junginio dėmeniu laikant pagalbinius žodžius (Osborne, Gerdes 2019: 2)

[84] "Universal Dependencies (UD) is a project that is developing cross-linguistically consistent treebank annotation for many languages" (52 interneto nuoroda).

[85] Prieiga internete: https://universaldependencies.org/introduction.html [žiūrėta 2022-11-22].

Būtent šiuo principu vaizduojami sakiniai lietuvių kalbos sintaksiškai anotuotame tekstyne ALKSNIS. Sakinys *Dalis požiūrių yra per mažai argumentuoti ir mums sunkiai priimtini*, pateiktas 81 pav., aiškiai parodo, kad jungtukas *ir* laikomas pagrindiniu dėmeniu (53 interneto nuoroda[86]). Tačiau tokiu atveju neparodomas ryšys tarp tiesiogiai sintaksiškai priklausomų žodžių ir sakinio sintezės metu gali trūkti informacijos valentingumui ar derinimui realizuoti.



81 pav. Sakinys, kuriame *ir* yra pagrindinis žodžių junginio dėmuo (53 interneto nuoroda)

4.2. Sintaksiniai analizatoriai

Lietuvoje buvo sukurti du sintaksiniai analizatoriai: vienas VDU Kompiuterinės lingvistikos centre, kitas – Vilniaus universitete. Vilniaus universiteto analizatorius naudojamas automatinio vertimo tikslams ir nėra viešai prieinamas. VDU sintaksinis analizatorius buvo parengtas projekto *Sematika-1*[87] metu. Jis gali perskaityti morfologiškai anotuotus tekstus ir pateikia keletą sintaksinės analizės variantų (Utka ir kt. 2016: 17).

[86] Prieiga internete: https://www.youtube.com/watch?v=PlE0PWurb4Y&t=29s 30-ta sekundė [žiūrėta 2022-11-22].

[87] Projektas Nr. VP2-3.1-IVPK-12-K-01-007.

4.2.1. Taisyklėmis pagrįstas metodas

Patys pirmieji sintaksiniai analizatoriai buvo kuriami naudojant taisyklių metodą, kuris yra labai imlus žmogaus darbui. 1978 m. Grenoblio universitete (Université Grenoble Alpes) sukurtoje automatinio vertimo sistemoje ARIANE-78, atliekant sintaksinę analizę, naudojamos 176 taisyklės (Guilbaud 1987: 297), o teoriniame darbe HARWARD SYNTACTIC ANALYSER anglų kalbai buvo sudaryta daugiau kaip 3 000 taisyklių (Winograd 1983: 360). Taisyklės sudaromos remiantis formaliomis gramatikomis.

4.2.1.1. Formalios gramatikos

Lingvistikos tyrinėtojams susidomėjus kalbų formalizavimu, buvo bandoma sudaryti matematinį kalbos modelį. Taip gimė formalios kalbos ir formalios gramatikos idėja. Formali kalba, kaip ir kiekviena kalba, turi savo abėcėlę, tik šios abėcėlės simbolius reikia suprasti abstrakčiau nei mums gerai žinomas 32 lietuviškas raides. Tai gali būti raidės, skaitmenys, skyrybos ženklai, net keli spausdinti ženklai gali būti vienas tos abėcėlės simbolis (Dagienė, Grigas 2007: 63). Formalios kalbos abėcėlę sudaro dviejų rūšių simboliai: galutiniai (angl. *terminal*) ir negalutiniai (angl. *nonterminal*). Galutiniai simboliai – tai kalbos, kurią reikia aprašyti (ar tai būtų kurios nors tautos kalba, ar programavimo kalba), vienetai. Tautų kalboms tai – morfologinės kategorijos ar tiesiog sakinio žodžiai. Negalutiniai simboliai nurodo pačios formalios kalbos sąvokas: tautų kalbose tai yra sintaksinės kategorijos, pvz., *daiktavardinė frazė*, *sakinys* ir kt. Galutinių simbolių seka vadinama eilute. Tautų kalbose sakinys yra žodžių seka (Batori, Lenders, Putschke 1989: 622). Taigi, žodžiai yra galutiniai simboliai, o sakinys – eilutė. Formalią kalbą sudaro eilučių rinkinys, tačiau ne visų galimų, o tik tam tikrų (taip pat kaip ir tautų kalbose: juk ne bet koks lietuviškų žodžių kratinys yra lietuvių kalbos sakinys). Ar eilutė priklauso formaliai kalbai, ar ne, nustatoma naudojantis formaliomis gramatikomis. Lygiai taip pat, kaip ir tautų kalbų atveju, pvz., pagal anglų kalbos gramatiką sakinys *Put some paper in the printer* yra galimas, o sakinys *Printer some put the paper in* – negalimas anglų kalboje (Arnold ir kt. 1994: 39). Kadangi formalios kalbos elementai yra eilutės, tai joms apdoroti taikomos eilučių operacijos. Programavimo kalbose naudojama eilučių keitimo komanda. Operacija $u \rightarrow v$ reiškia, kad simbolių seką u, rastą pradinėje eilutėje, reikia pakeisti

seka *v*. Taip gaunama nauja eilutė, pvz., jei turime eilutę *aab*, tai taikydami operaciją *ab* → *bav* gausime dvi naujas eilutes: *aab* → *abav* → *bavav*, o taikydami operaciją *b* → *bb* iš pradinės eilutės galime gauti begalinį skaičių naujų eilučių: *aab* → *aabb* → *aabbb* → *aabbbb*... Formalių gramatikų taisyklėms užrašyti ir buvo pasirinkta ši operacija, o pačios taisyklės vadinamos pakeitimo taisyklėmis.

Formalią gramatiką sudaro keturios dalys:

1) negalutinių (angl. *nonterminal*) simbolių baigtinė aibė N;
2) galutinių (angl. *terminal*) simbolių baigtinė aibė T;
3) gramatikos taisyklių baigtinė aibė P;
4) pradinis negalutinis simbolis S, nuo kurio pradedamas eilučių generavimas arba kuris turi būti gautas analizės metu.

Formalios gramatikos gali būti labai įvairios. Tą patį sakinį galima išnagrinėti pasitelkus kelias skirtingas gramatikas, t. y. kelis skirtingus pakeitimo taisyklių rinkinius. Kuri gramatika geriausia, nulemia trys kriterijai (Arnold ir kt. 1994: 44):

1) ar gramatikoje aprašyti visi kalbos sakiniai;
2) ar gramatika yra nepriekaištinga leistinų sakinių atžvilgiu, t. y. ar gramatiškai netaisyklingi sakiniai nelaikomi teisingais;
3) ar lengva ją suprasti ir taikyti atliekant automatinį kalbos apdorojimą.

4.2.1.2. Anglų kalbos sintaksinė analizė

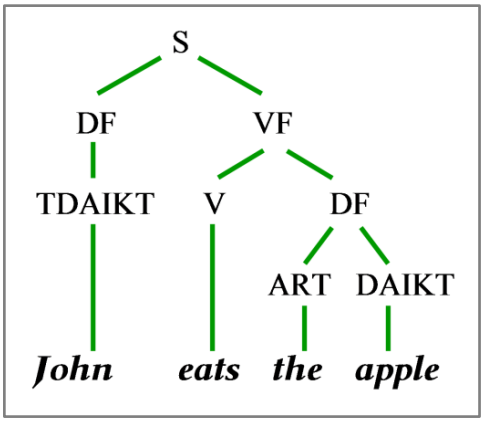
Anglų kalbos sakinį S paprastai sudaro daiktavardinė frazė DF ir veiksmažodinė frazė VF. Pastarąją sudaro veiksmažodis V ir daiktavardinė frazė DF. Sudėtingesnės veiksmažodinės frazės apima veiksmažodį V, daiktavardinę frazę DF ir prielinksninę frazę PF, t. y. prielinksninę konstrukciją.

Anglų kalbos sakiniams sintaksiškai išanalizuoti sudaromas pakeitimo taisyklių rinkinys, nusakantis, kokia sakinio struktūra yra galima. Kiekviena taisyklė nurodo, kad tam tikras simbolis medyje gali būti išskleistas kitų simbolių seka (Allen 1987: 41). Labai paprasto gramatikos taisyklių rinkinio pavyzdys pateikiamas 82 pav. Pirmos keturios taisyklės aprašo tik negalutinius simbolius, 5–8 taisyklės susieja negalutinius simbolius su galutiniais. Negalutiniai simboliai pažymėti didžiosiomis raidėmis, galutiniai – mažosiomis.

S –> DF VF (1)	TDAIKT –> John (5)
VF –> V DF (2)	V –> eats (6)
DF –> TDAIKT (3)	ART –> the (7)
DF –> ART DAIKT (4)	DAIKT –> apple (8)

82 pav. Pakeitimo taisyklių rinkinys (parengta pagal Allen 1987: 41)

Naudojant šias taisykles galima išnagrinėti tokį anglišką sakinį: *John eats the apple*. 83 pav. parodyta šio sakinio sintaksinė struktūra.



83 pav. Sakinio *John eats the apple* schema (parengta pagal Allen 1987: 41)

Nagrinėjant sakinį, pradedama nuo galutinių simbolių eilutės, t. y. nuo konkretaus sakinio, ir žodžiai keičiami jų sintaksinėmis kategorijomis. 84 pav. parodyta šio sakinio nagrinėjimo, t. y. kompiuterinės analizės, eiga.

-> John eats the apple –>	(pagal 5 taisyklę)
–> TDAIKT eats the apple –>	(pagal 6 taisyklę)
–> TDAIKT V the apple –>	(pagal 7 taisyklę)
–> TDAIKT V ART apple –>	(pagal 8 taisyklę)
–> TDAIKT V ART DAIKT –>	(pagal 3 taisyklę)
–> DF V ART DAIKT –>	(pagal 4 taisyklę)
–> DF V DF –>	(pagal 2 taisyklę)
–> DF VF–>	(pagal 1 taisyklę)
–> S	

84 pav. Sakinio *John eats the apple* analizė (parengta pagal Allen 1987: 41)

Šiuo atveju pakeitimo taisyklių dešinėje rodyklės pusėje esantys simboliai (82 pav.) keičiami kairės pusės simboliais. Jei taikant taisykles pavyksta gauti pradinį simbolį S, vadinasi, sakinys yra gramatiškai taisyklingas. Reiktų pasakyti, kad sintaksės požiūriu sakinio reikšmė yra visai neaktualu. Sakinys turi būti išnagrinėtas formaliai, visiškai nesigilinant į jo prasmę. Todėl net ir logiškai neteisingam sakiniui kompiuteris privalo sugebėti sudaryti sintaksinę struktūrą, jei tik jis formaliai, gramatiškai yra taisyklingas.

Naudojama ir kita, atvirkštinė aprašytai, metodika, kai analizė pradedama nuo sakinio, t. y. nuo pradinio simbolio, ir, taikant taisykles, jų kairėje pusėje esantys simboliai keičiami dešinės pusės simboliais. Keitimai atliekami tol, kol nebelieka negalutinių simbolių, t. y. kol gaunamas sakinys.

Sintaksinės analizės pobūdis gali būti labai skirtingas, ir tai priklauso nuo kalbos. Rusų kalboje daug ką lemia daiktavardžio bei būdvardžio linksniai ir veiksmažodžio formos. Šiek tiek informacijos gali turėti ir žodžių tvarka. Anglų kalboje beveik viską nulemia žodžių tvarka, o daiktavardžio ir veiksmažodžio formų variantai atliekant sintaksinę analizę mažai tegali padėti (Henisz-Dostert, Macdonald, Zarechnak 1979: 116). Pagal galūnes anglų kalboje sakinio dalys nustatomos tik labai retais atvejais, pvz., *I know Danny and Toni know**s** me* ir *I know Danny and Toni know me* (Winograd 1983: 136). Taigi, ir lietuvių kalboje, „neturinčioje griežtos sugramatintos žodžių tvarkos“ (Labutis 2002: 15), negalima tikėtis gerų rezultatų, taikant anglų kalbai naudojamas metodikas.

4.2.1.3. Lenkų kalbos sintaksinė analizė

Lenkų kalbos sintaksinio analizatoriaus sukūrimo tikslas buvo panaudoti jį ruošiant ontologiją, tais atvejais, kai tenka automatiškai surinkti duomenis (Adamski, Zimniewicz 2011: 536). Sintaksinės analizės algoritmas (detali veiksmų seka) turi šešis etapus:

1) teksto segmentavimas (suskirstymas sakiniais) ir morfologinė analizė,
2) žodžių grupavimas,
3) sudėtinių sakinių išskaidymas į vientisinius,
4) veiksnio ir tarinio atpažinimas,
5) papildinių atpažinimas pagal valentingumą,
6) konteksto analizė.

Pirmąjį etapą sudaro trys žingsniai: teksto suskaidymas į sakinius, nedalomų vienetų (žodžių) suradimas ir morfologinė analizė. Ji atliekama panaudojant Vroclavo technikos universitete (Politechnika Wrocławska) sukurtą programinę įrangą: nustatoma žodžio pradinė forma, kalbos dalis, giminė, skaičius, linksnis ir asmuo. Kadangi žodžių formų atpažinimas kartais gali būti daugiareikšmis, todėl šio žingsnio tikslumas turi labai didelę įtaką visos tolimesnės analizės rezultatams.

Antrojo etapo metu žodžiai, susiję su tuo pačiu objektu, grupuojami į stambesnius vienetus. Pavyzdžiui, sakinyje *W drugiej publikacji zostal podjęty temat roli Europejskiego Banku Centralnego* išskiriamos dvi grupės: *w drugiej publikacji* ir *Europejskiego Banku Centralnego*. Tikslas – supaprastinti tolimesnę sintaksinę analizę, sumažinant nepriklausomų vienetų skaičių sakinyje. Iš pradžių grupuojant ieškoma šalia esančių žodžių, kurie būtų to paties linksnio, skaičiaus ir giminės, pvz., *Europejskiego Banku Centralnego* visi trys žodžiai yra vyriškosios giminės vienaskaitos kilmininkas. Toliau žodžiai grupuojami, jei po prielinksnio einantis žodis (ar žodžiai) yra vietininko linksnio ir tokia žodžių grupė laikoma aplinkybe, nes ji dažniausiai aprašo vietą ar laiką, pvz., *w drugiej publikacji*.

Trečiajame etape sudėtiniai sakiniai skaidomi į vientisinius ir atliekama jų sintaksinė analizė. Sudėtiniai sakiniai nustatomi analizuojant galimus skaidymo taškus, pvz., jei randamas jungtukas arba kablelis ir abiejose jų pusėse yra po veiksnį ir tarinį.

Kitame etape ieškoma veiksnio ir tarinio. Tarinio funkcija paprastai priskiriama asmenuojamajai veiksmažodžio formai, o veiksnio funkcija – vardininko linksnio daiktavardžiams ar įvardžiams. Šio etapo metu surandami ir prieveiksmiai, kuriems priskiriama aplinkybės funkcija.

Papildiniai, išplečiantys tarinį, nustatomi jau vėlesniame etape. Jų ieškoma naudojantis lenkų kalbos valentingumo žodynu (Przepiorkowsky 2008), kuris apima 2 000 veiksmažodžių ir tai sudaro daugiau nei 90 proc. dažniausiai vartojamų žodžių. Jei nagrinėjamas veiksmažodis nerandamas valentingumo žodyne, ieškoma galininko ir kilmininko linksnio žodžių. Netiesioginių papildinių, kurie dažniausiai būna naudininko linksnio, ieškoma atskirai.

Paskutinis etapas yra skirtas antecedentams nustatyti sudėtiniuose sakiniuose, pvz., *który jest zielony*. Nors šis žingsnis vadinamas konteksto analize, tačiau apsiribojama tik sudėtiniuose sakiniuose esančiais žodžiais. Jei šalutiniame sakinyje aptinkamas jungiamasis žodis, pvz., *który*, prieš jį esančiame sakinyje ieškoma paskutinio žodžio, kuris yra tos pačios giminės ir skaičiaus. Tačiau toks priskyrimas

gali būti ir klaidingas, jei prieš tai esančiame sakinyje yra ne vienas žodis, sutampantis gimine ir skaičiumi. Čia jau reikėtų papildomos informacijos valentingumo žodyne apie tai, koks papildinys gali būti tam tikro veiksmo atlikėju, pvz., gyvas ar negyvas. Toks lenkų kalbos žodynas yra parašytas, bet kol kas jis išleistas tik kaip knyga.

Apibendrinant galima pasakyti, kad lenkų sintaksinė analizė kuriama daugiausia atsižvelgiant į tai, kaip sakinį nagrinėja žmogus (Adamski, Zimniewicz 2011: 538).

4.2.1.4. Lietuvių kalbos automatinė sintaksinė analizė

Lietuvių kalbos sintaksinis analizatorius, sukurtas naudojant taisyklėmis pagrįstą metodą (Boizou, Zamblera 2014: 70), yra *Lietuvių kalbos sintaksinės ir semantinės analizės informacinės sistemos* dalis. Programinė įranga sudaryta iš keturių modulių:

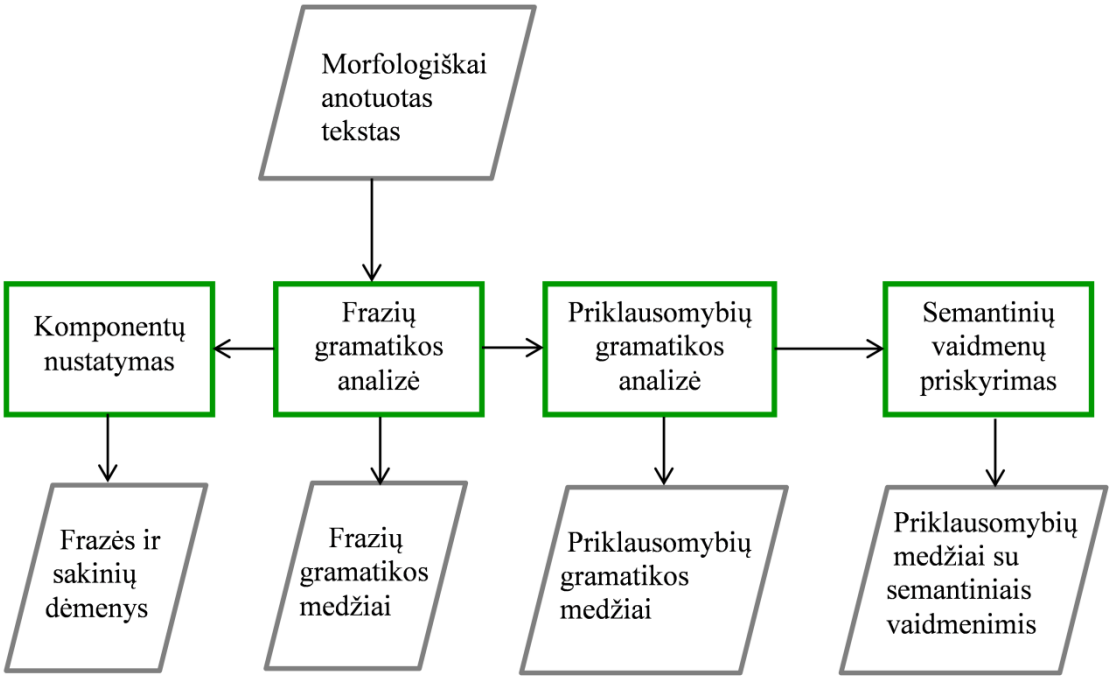
1) sakinio analizė, atliekama frazių gramatikos principu,
2) sakinio komponentų nustatymas,
3) sakinio analizė, atliekama naudojant priklausomybių gramatikos metodą,
4) priklausomybių gramatikos papildymas semantiniais vaidmenimis.

Iš esmės pats sintaksinis analizatorius tiktai sudaro sakinio struktūrą frazių gramatikos metodu. Jam turi būti pateikiami morfologiškai anotuoti sakiniai, kuriuose jau panaikintas daugiareikšmiškumas. Visi trys kiti moduliai kaip pradinius duomenis ima frazių gramatikos modulio darbo rezultatus ir apdoroja juos pagal savo paskirtį: formuoja priklausomybių medį ir kt. 85 pav. parodyta analizatoriaus struktūrinė schema.

Visi keturi moduliai naudojasi ta pačia gramatika. Ji, kaip ir kiekviena formali gramatika, turi negalutinių simbolių rinkinį, pakeitimo taisyklių rinkinį ir pradinį simbolį. Galutiniai simboliai, t. y. žodžiai ir jų morfologiniai duomenys, nėra išvardyti taisyklių rinkinyje. Bendra visiems moduliams gramatika buvo papildyta dviejų tipų informacija:

a) nurodytas pagrindinis žodžių junginio dėmuo,
b) pakeitimo taisyklės suskirstytos į tris lygmenis: pagrindinės, antraeilės ir retos.

Taip papildžius gramatiką atsiranda galimybė pateikti vieną analizuojamo sakinio sintaksinį medį vietoje kelių alternatyvių jo variantų.



85 pav. Lietuvių kalbos sintaksinio analizatoriaus struktūra (parengta pagal Boizou, Zamblera 2014: 70)

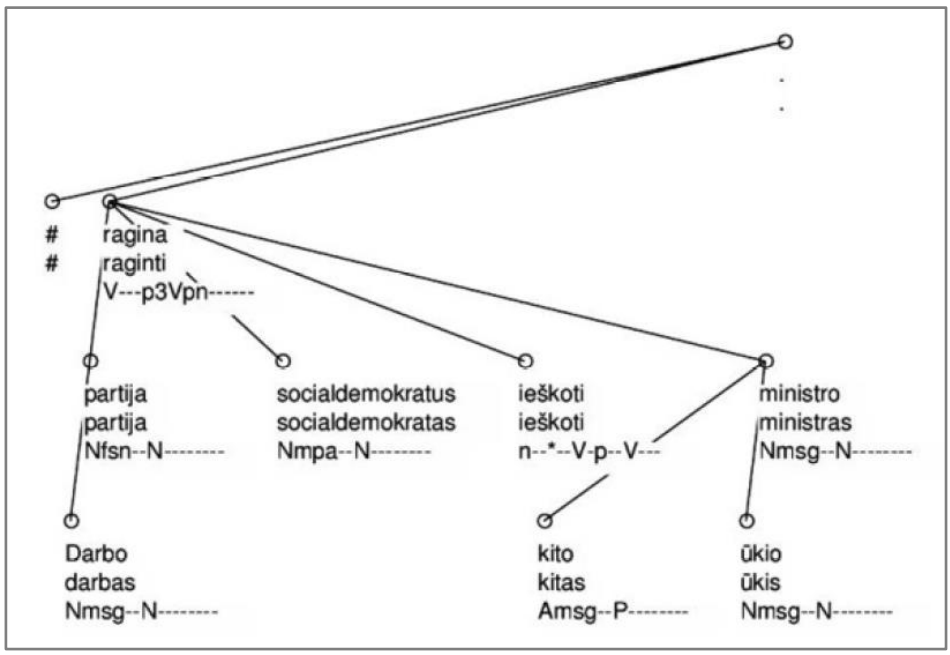
Turint išanalizuotą frazių gramatikos metodu sakinį, antrajame modulyje išskiriami jo komponentai, t. y. nustatomos frazės (sintagmos) ir sudėtinio sakinio dėmenys. Rezultatas – visų dėmenų sąrašas, pateikiant pagrindinius sakinius su jų viduje esančiais prijungiamaisiais sakiniais. Pavyzdžiui, sakinys *Krūmas – dažnai tavo ūgio ar kiek didesnis, jį sudaro daug sumedėjusių stiebų, kuriuos krūmai išleidžia iš kelmo* išskaidytas į komponentus atrodys taip:

1) *Krūmas – dažnai tavo ūgio ar kiek didesnis,*
2) *jį sudaro daug sumedėjusių stiebų, kuriuos krūmai išleidžia iš kelmo,*
3) *kuriuos krūmai išleidžia iš kelmo.*

Nesunku nustatyti daiktavardines ir prielinksnines frazes, tačiau su veiksmažodinėmis frazėmis iškyla problemų dėl laisvos žodžių tvarkos lietuvių kalbos sakiniuose, todėl jose paprastai apsiribojama veiksmažodžio ir prieveiksmio junginiu. Trečiasis pateikto sakinio komponentas išskaidomas į tris frazes: daiktavardinė frazė – *krūmai*, veiksmažodinė frazė – *išleidžia* ir prielinksninė frazė – *iš kelmo* (Boizou, Zamblera 2014: 72).

Trečiasis modulis iš frazių gramatikos medžio suformuoja priklausomybių gramatikos medį. 86 pav. pateiktas sakinio *Darbo partija ragina socialdemokratus ieškoti kito ūkio ministro* priklausomybių medis. Nors patys autoriai teigia, kad pakeisti frazių

gramatikos medį į priklausomybių gramatikos medį yra trivialu[88] (Boizou, Zamblera 2014: 72), t. y. labai paprasta, elementaru, bet 86 pav. matyti, kad tarp žodžių *ragina* ir *ministro* yra tiesioginis ryšys. Tačiau žodis *ministro* išplečia žodį *ieškoti*, o ne žodį *ragina*, t. y. tiesioginiu sintaksiniu ryšiu yra susiję žodžiai *ieškoti ministro*, o ne *ragina ministro*. Gaila, kad nepateikiamas šio sakinio frazių gramatikos medis, taigi, negalima matyti, kurioje vietoje atsiranda klaida.



86 pav. Sakinio *Darbo partija ragina socialdemokratus ieškoti kito ūkio ministro* priklausomybių medis (parengta pagal Boizou, Zamblera 2014: 72)

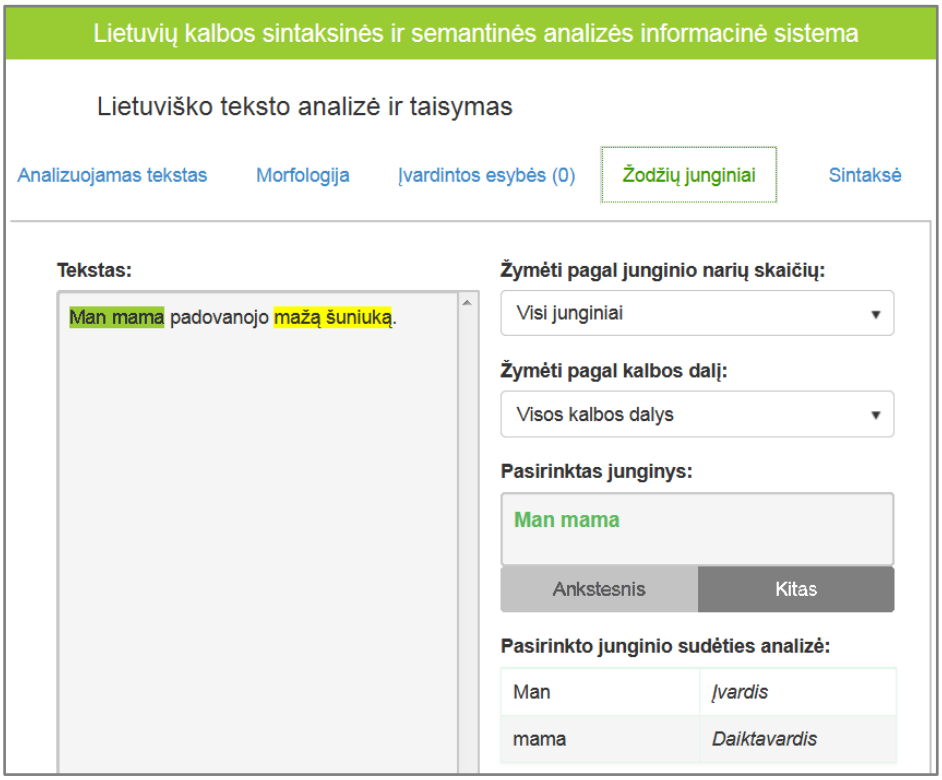
Lietuvių kalbos sintaksinio analizatoriaus priklausomybių medyje nėra pateikiama sintaksinė informacija, pvz.: veiksnys, pažyminys ir kt. Toks pasirinkimas grindžiamas faktu, kad skirtumai tarp papildinio ir aplinkybės yra susiję su semantiniais vaidmenimis, todėl, siekiant išvengti informacijos pasikartojimo, šie duomenys paliekami ketvirtajam moduliui, kuriame semantiniai vaidmenys žodžiams priskiriami naudojantis Nijolės Sližienės žodyne (Sližienė 1994–2005) pateiktu principu.

VDU *Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema* buvo prieinama laisvai internete iki 2020 m. vasario 21 d. (42 interneto nuoroda[89])

[88] "[...] dependency derivation is a trivial task" (Boizou, Zamblera 2014: 72).

[89] Prieiga internete: https://web.archive.org/web/20200221090527/http://www.semantika.lt:80/SyntaticAndSemanticAnalysis/Analysis [žiūrėta 2022-11-22].

Sintaksinė informacija vartotojui buvo pateikiama nurodant žodžių junginius. Sakiniui *Man mama padovanojo mažą šuniuką* nustatyti žodžių junginiai parodyti 87 pav. (43 interneto nuoroda[90]).



87 pav. Žodžių junginių nustatymas sakinyje

Kiekvienam surasto žodžių junginio žodžiui buvo nurodoma kalbos dalis. Žodžių junginys *mažą šuniuką* nustatytas teisingai, tačiau žodžiai *man* ir *mama*, kurie pateikti kaip žodžių junginys, tiesioginio sintaksinio ryšio neturi. Šiuo metu atnaujinta sintaksinės dalies versija tinklalapyje *semantika.lt* kol kas nepateikta.

4.2.2. Statistiniai metodai

Statistiniai sintaksiniai analizatoriai susieja formalių gramatikų taisykles su tikimybėmis. Jie sudaro visus galimus sakinio analizės variantus ir tada apskaičiuoja kiekvieno jų tikimybę. Kaip rezultatas pateikiamas labiausiai tikėtinas variantas[91] (Charniak 1997: 37).

[90] Prieiga internete: www.semantika.lt/SyntaticAndSemanticAnalysis/Analysis [žiūrėta 2019-12-20].

[91] "Statistical parsers work by assigning probabilities to possible parses of a sentence, locating the most probable parse, and then presenting the parse as the answer" (Charniak 1997: 37).

Galima tai pailiustruoti pavyzdžiu – sakinio *The can can hold the water* analize. Čia iš karto matome daiktą, vadinamą *the can*, ir šis daiktas gali atlikti veiksmą *can* (t. y. sugebėti), ir tai, ką šis daiktas sugeba daryti, yra *hold*, kurio objektas – *water*. Tačiau tai nėra vienintelis galimas šio sakinio analizės variantas (54 interneto nuoroda[92]). *The can can* yra miuzikholo šokio pavadinimas, kuris buvo populiarus apie 1840 m. ir toks išliko iki šių dienų prancūzų kabaretuose (55 interneto nuoroda[93]), o *hold water* taip pat yra leistina veiksmažodinė frazė. Tai dar vienas formaliai galimas analizės variantas, tačiau tokio sakinio reikšmė nėra labai aiški ir akivaizdi. Todėl pirmuoju atveju gauta sakinio analizė yra labiau priimtina ir statistiniai analizatoriai tai nusprendžia, surikiuodami gautus sakinio analizės variantus pagal jų tikimybes (54 interneto nuoroda).

Naujausi statistiniai analizatoriai remiasi automatiniu mokymusi iš tekstynų, kurie jau yra sintaksiškai anotuoti žmogaus, todėl sistema gali išgauti informaciją apie tai, su kokia tikimybe įvairios konstrukcijos pasitaiko tam tikrame kontekste (56 interneto nuoroda[94]).

4.2.2.1. Latvių kalbos sintaksinis analizatorius

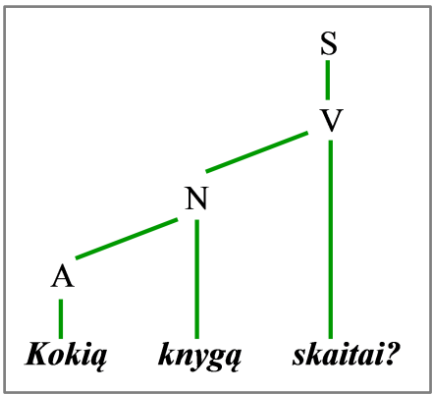
Statistiniais metodais veikiantis latvių kalbos sintaksinis analizatorius buvo kuriamas naudojant sintaksiškai anotuotus tekstus pagal *Prahos sintaksiškai anotuoto tekstyno* pavyzdį (žr. 37 pav.), tačiau jame sukauptos tik trijų lygmenų žymos. Ketvirtasis – tektogramatinis – lygmuo, nurodantis semantinę informaciją, nebuvo įtrauktas (Pretkalniņa, Rituma 2013: 280). Apmokymo duomenis sudaro apie 2 500 ranka morfologiškai anotuotų sakinių. Kad būtų galima patyrinėti klaidas, atsirandančias dėl neteisingo morfologinio anotavimo, buvo atliktas atskiras eksperimentas, kurio metu tekstai buvo morfologiškai anotuoti naudojant statistiniais metodais veikiantį morfologinį anotatorių.

Atlikta keletas eksperimentų, taikant įvairius sintaksinės analizės algoritmus ir jiems pateikiant įvairius morfologinių duomenų rinkinius. Kadangi *Latvių kalbos sintaksiškai anotuotas tekstynas* (angl. *Latvian Treebank*, latv. *Latviešu valodas sintaktiski marķētais korpuss*) sudarytas neprojektyvaus anotavimo metodu, dalis

[92] Prieiga internete: https://en.wikipedia.org/wiki/Statistical_parsing [žiūrėta 2022-11-22].
[93] Prieiga internete: https://en.wikipedia.org/wiki/Can-can [žiūrėta 2022-11-22].
[94] Prieiga internete: https://en.wikipedia.org/wiki/Parsing [žiūrėta 2022-11-22].

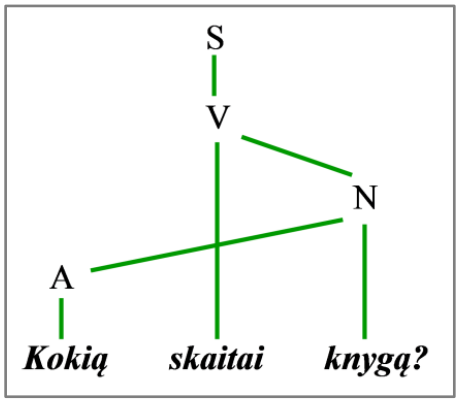
eksperimentų buvo atlikta pasitelkus neprojektyvių duomenų apdorojimo algoritmą (Pretkalniņa, Rituma 2013: 284). Čia reikėtų kiek išsamiau aptarti terminus *projektyvus* (angl. *projective*) ir *neprojektyvus* (angl. *non-projective*).

Projektyvūs ir neprojektyvūs sakiniai. Daugumą sakinių priklausomybių gramatikoje galima suprojektuoti į linijinę struktūrą taip, kad nebūtų besikryžiuojančių linijų (Holvoet 2009: 24). Tokie sakiniai vadinami projektyviais, pvz., sakinys *Kokią knygą skaitai?* yra projektyvus (88 pav.).



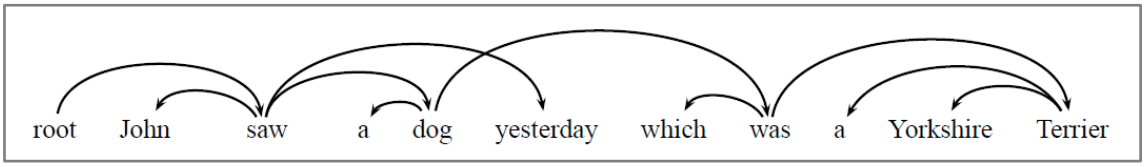
88 pav. Projektyvus sakinys (parengta pagal Holvoet 2009: 25)

89 pav. pateiktas neprojektyvaus sakinio pavyzdys: žodžių rinkinys lieka tas pats, keičiasi tik žodžių tvarka sakinyje, ir atsiranda linijų susikirtimas.



89 pav. Neprojektyvus sakinys (parengta pagal Holvoet 2009: 26)

Ypač dažnai neprojektyvumo reiškinys pastebimas sudėtiniuose prijungiamuosiuose sakiniuose, jų pasitaiko net ir anglų kalboje. 90 pav. pateiktas anglų kalbos sakinio pavyzdys, kuriame irgi matomos besikertančios linijos (McDonald ir kt. 2005: 2).



90 pav. Neprojektyvus anglų kalbos sakinys (McDonald ir kt. 2005: 2)

Siekiant išsamiau ištirti latvių kalbos sintaksinio analizatoriaus veikimą, buvo atlikta keletas eksperimentų, naudojant įvairius morfologinių duomenų rinkinius. Morfologinių žymų rinkinys apima 500 vienetų. Iš pradžių buvo atliktas testavimas naudojant jas visas, t. y. su išsamiu žodžių aprašymu, o vėliau buvo testuojama panaikinus dalį žymų, pvz., leksines žymas: tikrinis ar bendrinis daiktavardis, linksniuotė ir kt. (Paikens, Rituma, Pretkalniņa 2013: 269). Paskutiniame eksperimento etape buvo suskaidytos morfologinių žymų sekos, kiekvieną morfologinį požymį pateikiant atskirai, kad būtų sudarytos galimybės analizatoriui išmokti tokius dėsningumus kaip „nuo daiktavardžio priklausomas būdvardis turi būti suderintas su juo gimine, skaičiumi ir linksniu“ ir pan. (Pretkalniņa, Rituma 2013: 284).

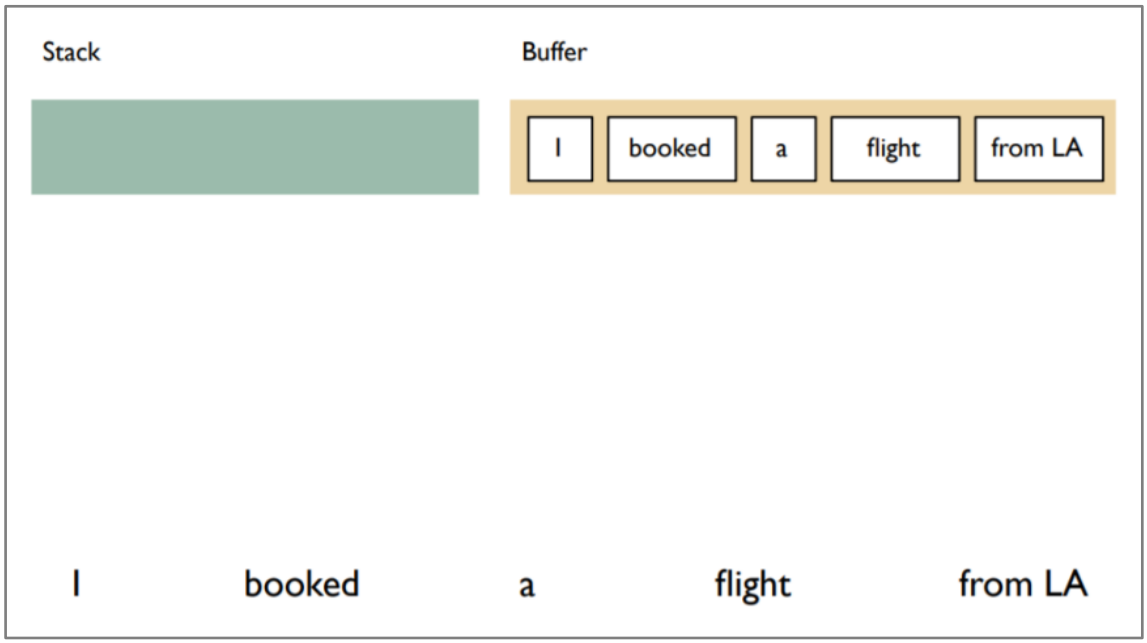
Latvių kalbos sintaksinio analizatoriaus tikslumas buvo 74,63 proc., naudojant ranka morfologiškai anotuotus sakinius, ir jis nukrito iki 72,2 proc., kai morfologiškai tekstynas buvo anotuotas automatiniu būdu (Pretkalniņa, Rituma 2013: 286).

4.2.2.2. Lietuvių kalbos sintaksiniai analizatoriai

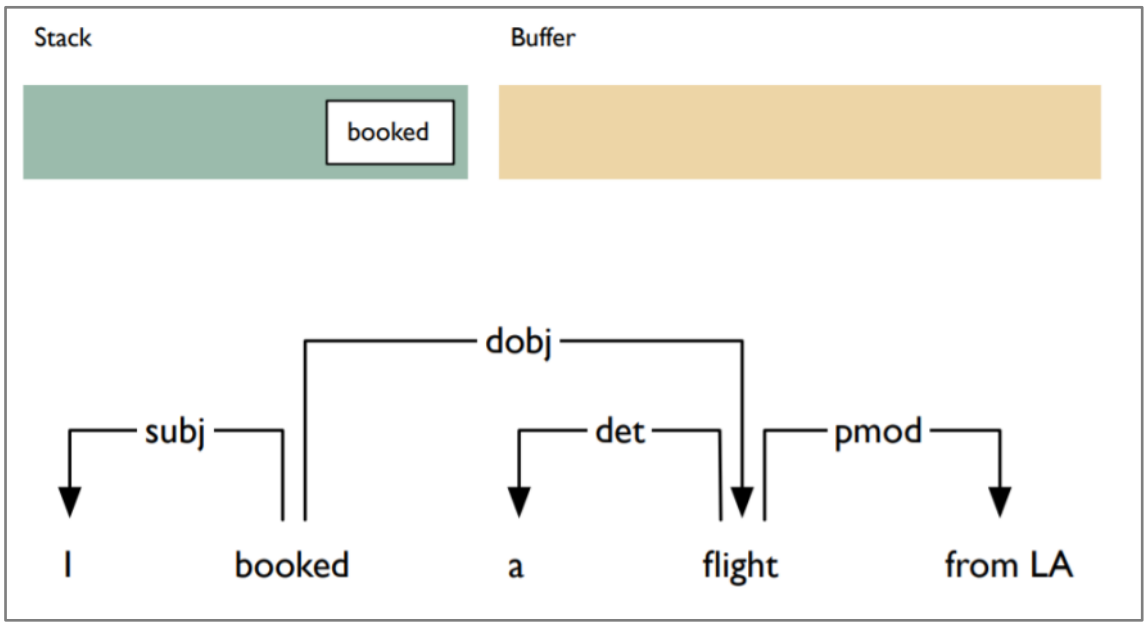
Švedijoje sukurto statistiniais metodais veikiančio sintaksinio analizatoriaus *MaltParser* veikimas paremtas apmokymu iš sintaksiškai anotuotų tekstynų (Nivre ir kt. 2007: 96). Sintaksinei analizei atlikti jame taikomas būsenų kaitos metodas (angl. *transition-based parser*). Verta išsamiau aprašyti šį metodą.

Būsenų kaitos metodas. *MaltParser* analizatoriuje sintaksinė analizė atliekama būsenų kaitos (angl. *transition-based*) metodu, naudojant lankų algoritmą (Stymne 2014: 9). Analizės metu sudaromas tik vienas priklausomybių medis ir tik vieną kartą peržiūrimas sakinys iš kairės į dešinę. Žinoma, nėra garantijos, kad bus suformuota pati tiksliausia sakinio sintaksinė struktūra (Stymne 2014: 34). Sakiniui apdoroti reikia trijų komponentų: buferio, steko (tai – laikinoji informacijos saugykla) ir priklausomybių medžio. Analizės pradžioje sakinio žodžiai sukeliami į buferį, stekas yra tuščias, o priklausomybių medyje surašyti tik sakinio žodžiai (91 pav.). Baigus analizę, buferis turi būti tuščias, steke turi būti likęs tik vienas žodis – medžio šaknis, o

priklausomybių medyje – linijomis parodyti ryšiai tarp žodžių (92 pav.). Pagrindiniai atliekami veiksmai: žodžio perkėlimas iš buferio į steką, lanko į kairę suformavimas ir lanko į dešinę suformavimas. Kaip suformuoti lankus, analizatorius apmokomas naudojant auksinį standartą (Stymne 2014: 34). Detali angliško sakinio *I booked a flight from LA* analizė parodyta 7 priede.



91 pav. Lankų algoritmo pradinė būsena (Stymne 2014: 14)



92 pav. Lankų algoritmo galinė būsena (Stymne 2014: 33)

MaltParser buvo sėkmingai pritaikytas lietuvių kalbai (Kapočiūtė-Dzikienė, Damaševičius 2020: 455). Pasinaudojant sintaksiškai anotuotu lietuvių kalbos tekstynu ALKSNIS, buvo atliekami tyrimai, kurių tikslas – išsiaiškinti, kokie žodžių morfologiniai požymiai turi didžiausią įtaką sintaksinės analizės rezultatams. Sintaksinio analizatoriaus darbo rezultatai buvo vertinami dviem aspektais: ryšių tarp žodžių suradimo (angl. *unlabelled attachment score* – UAS) ir ryšių bei žymų nustatymo (angl. *labelled attachment score* – LAS). 93 pav. parodyta, kaip buvo skaičiuojami įverčiai. Geriausi pasiekti rezultatai: 76,2 proc. (UAS) ir 71,9 proc. (LAS).

$$UAS = \frac{teisingai\ nustatytos\ priklausomybės}{visos\ priklausomybės} \times 100\ \%$$

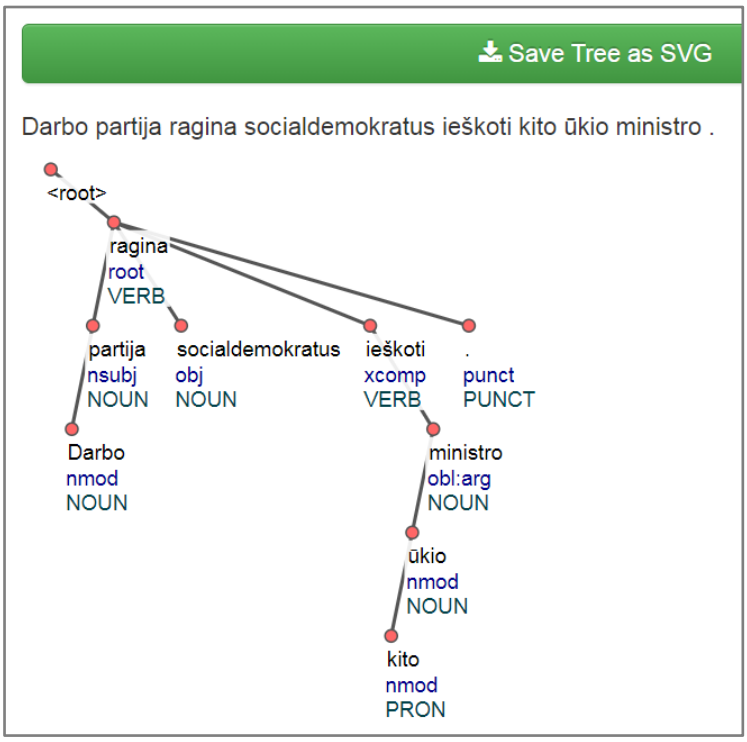
$$LAS = \frac{teisingai\ nustatytos\ priklausomybės\ ir\ žymos}{visos\ priklausomybės\ ir\ žymos} \times 100\ \%$$

93 pav. *MaltParser* įverčių apskaičiavimas (parengta pagal Kapočiūtė-Dzikienė, Damaševičius 2020: 456)

Sakinių analizės pavyzdžiai. Statistiniais metodais veikiančio sintaksinio anaizatoriaus *UDPipe* lietuvių kalbos modulis yra laisvai prieinamas internete (45 interneto nuoroda[95]). Apmokymui buvo naudotas lietuvių kalbos sintaksiškai anotuotas tekstynas ALKSNIS. Vartotojo įvestam sakiniui analizatorius pateikia trijų tipų analizės rezultatus: CoNLL-U failą, lentelę ir priklausomybių medį. Palyginimui galima paimti tą patį sakinį, kuris buvo aprašytas taisyklėmis pagrįsto sintaksinio analizatoriaus skyriuje (4.2.1.4 poskyryje). Sakinio *Darbo partija ragina socialdemokratus ieškoti kito ūkio ministro* morfologinė analizė atlikta teisingai. Sintaksiniai ryšiai tik vienam žodžiui nustatyti netiksliai: žodis *kito* parodytas kaip susijęs su žodžiu ūkio, nors teisingas jo ryšys būtų su žodžiu *ministro*, kuris liko neparodytas (94 pav.).

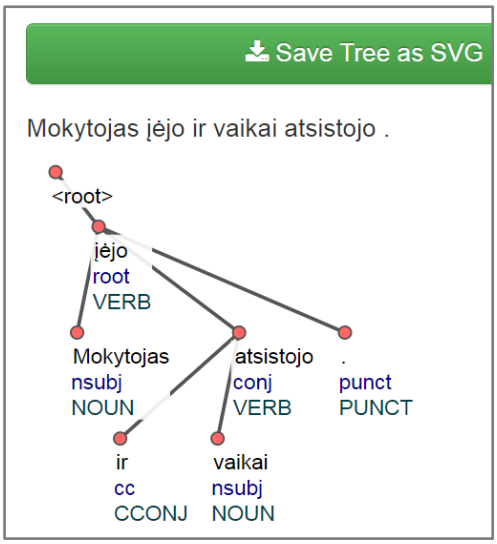
Taisyklių metodu veikiantis analizatorius šioje vietoje klaidos nepadarė. Ten buvo kita klaida: neteisingai susieti žodžiai *ragina* ir *ministro*, o tarp žodžių *ragina* ir *ieškoti* ryšys nenustatytas (86 pav.)

[95] Prieiga internete: https://lindat.mff.cuni.cz/services/udpipe/run.php [žiūrėta 2022-11-22].



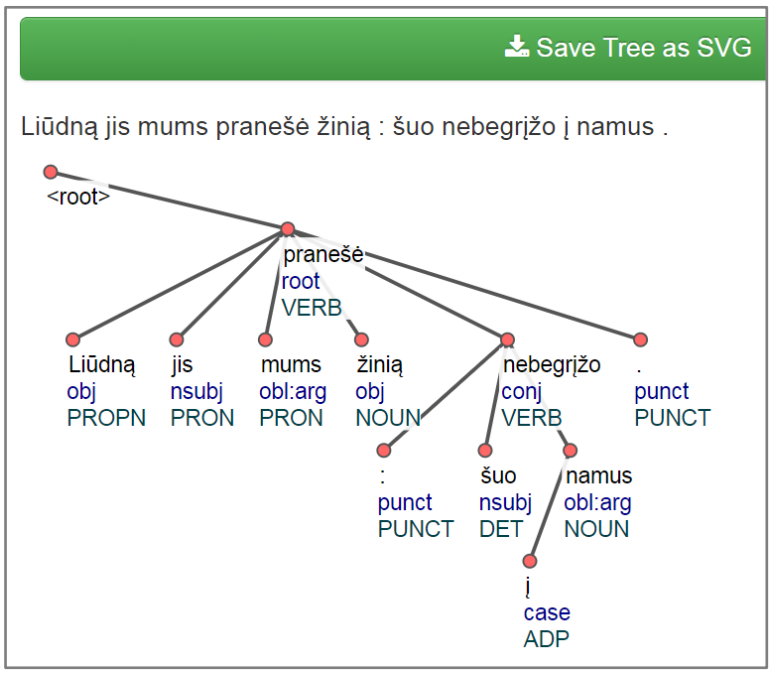
94 pav. Sakinio *Darbo partija ragina socialdemokratus ieškoti kito ūkio ministro* priklausomybių medis (*UDPipe* 45 interneto nuoroda)

Kitas pavyzdys galėtų būti 6 priede pateikto sakinio *Mokytojas įėjo ir vaikai atsistojo* sintaksinė analizė. Nors abu sudėtinio sujungiamojo sakinio dėmenys yra lygiaverčiai, schemoje antrojo dėmens tarinys vaizduojamas kaip priklausomas nuo pirmojo (95 pav.).



95 pav. Sakinio *Mokytojas įėjo ir vaikai atsistojo* priklausomybių medis (*UDPipe* 45 interneto nuoroda)

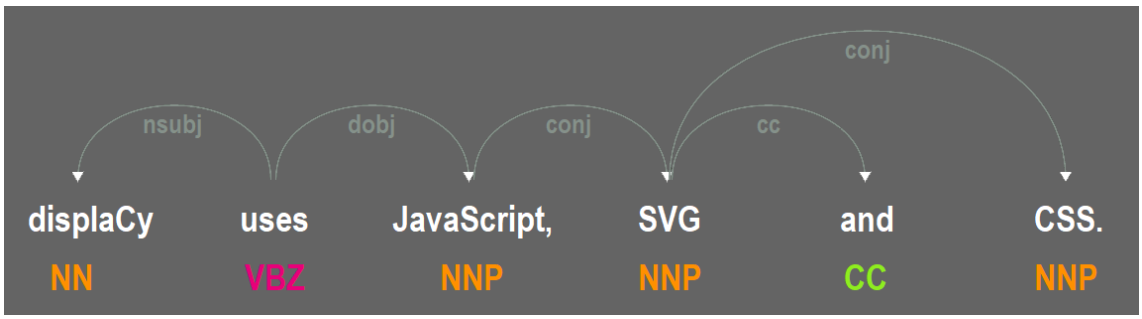
Morfologijos dalyje aptarto sakinio *Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus* (3.1.2.1 poskyris) sintaksinė struktūra atrodo taip (96 pav.):



96 pav. Sakinio *Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus* priklausomybių medis (*UDPipe* 45 interneto nuoroda)

Klaidos morfologinės analizės metu nulėmė ir sintaksinės struktūros netikslumus. Kadangi žodžiui *Liūdną* buvo suteiktas požymis *tikrinis daiktavardis* (51 pav.), todėl sintaksinėje struktūroje jis yra papildinys, t. y. toks pat kaip ir *žinią*. Lieka neparodytas tiesioginis ryšys tarp žodžių *liūdną* ir *žinią*.

Dar vienas analizatorius, kurio apmokymui buvo naudotas lietuvių kalbos sintaksiškai anotuotas tekstynas ALKSNIS, yra atvirojo kodo programinė įranga *SpaCy* (57 interneto nuoroda[96]). Ji buvo sukurta 2016 m. ir pateikia sakinio struktūrą grafo lankų metodu (97 pav.).



97 pav. *SpaCy* pateikiamo sakinio pavyzdys (57 interneto nuoroda)

[96] Prieiga internete: https://en.wikipedia.org/wiki/SpaCy [žiūrėta 2022-11-22].

2021 m. pasirodė trečioji versija – *SpaCy 3.0* – ir buvo parengtas lietuvių kalbos modelis. Internete pateikiamas lietuviško sakinio *Jaunikis pirmąją vestuvinę naktį iškeitė į areštinės gultą* analizės pavyzdys (58 interneto nuoroda[97]), jis parodytas 98 pav. Galimi trys modelio variantai: siauros apimties, vidutinės ir didelės apimties. Siauros apimties modelyje pateikiama mažiau informacijos: tik kalbos dalis ir sintaksinė funkcija, būtent šio modelio pavyzdys parodytas 98 pav.

```
import spacy
from spacy.lang.lt.examples import sentences

nlp = spacy.load("lt_core_news_sm")
doc = nlp(sentences[0])
print(doc.text)
for token in doc:
    print(token.text, token.pos_, token.dep_)
```

RUN

```
Jaunikis pirmąją vestuvinę naktį iškeitė į areštinės gultą
Jaunikis PROPN nsubj
pirmąją NUM nummod
vestuvinę ADJ amod
naktį NOUN obj
iškeitė NOUN ROOT
į ADP case
areštinės ADJ amod
gultą NOUN obl:arg
```

98 pav. Lietuvių kalbos sakinio analizė, atlikta su *SpaCy 3.0*: siauros apimties modelis (58 interneto nuoroda)

Bendriniam daiktavardžiui *jaunikis* nurodytas požymis *tikrinis daiktavardis* turbūt dėl tos pačios priežasties kaip ir žodžiui *liūdną* (žr. 96 pav.) analizatoriumi *UDPipe* atliktoje sakinio *Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus*

[97] Prieiga internete: https://spacy.io/models/lt [žiūrėta 2022-11-22].

analizėje – daiktavardis parašytas didžiąja raide. Taigi, neįvertinama, kad tai žodis, kuriuo prasideda sakinys. Veiksmažodžiui *iškeitė* suteiktas kalbos dalies požymis – *daiktavardis*, toks pat kaip ir žodžiams *naktį* ir *gultą*. Būtent šioje vietoje labai gerai matyti esminis neuroninių tinklų metodo trūkumas – niekas negali pasakyti, kodėl žodžiui *iškeitė* neteisingai nustatyta kalbos dalis, ir svarbiausia, neaišku, ką reikia daryti, kad tos klaidos neliktų, nes net ir pateikus daugiau apmokymo duomenų, negalime būti tikri, kad tokios klaidos nebepasikartos. Tradicinio (taisyklėmis grįsto) programavimo atveju, analizuojant programos kodą visada galima rasti, kurioje jo vietoje atsiranda klaida, ir pakeitus tą kodo dalį, t. y. ištaisius programos kodą, tokio tipo klaidos niekada nebebus daromos. Bet tradicinis programavimas yra labai imlus darbui, todėl pasirenkamas greičiau sukuriamas variantas – neuroniniai tinklai, prarandant kokybę, t. y. tikslumą.

Didelės apimties modelyje nurodoma dar ir žodžio pradinė forma bei morfologinės kategorijos. 99 pav. pateikiamas sakinio *Liūdną jis mus pranešė žinią: šuo nebegrįžo į namus* analizė. Šis analizatorius nepadarė klaidų nustatydamas žodžio *šuo* lemą ir morfologines kategorijas, kurios buvo analizuojant šį sakinį su *UDPipe*. Tačiau įvardžiui *mums* lema su *SpaCy* nustayta neteisingai, kaip ir veiksmažodžiui *nebegrįžo*. Kitą klaidą lėmė, matyt, didelis atstumas tarp tiesioginiu sintaksiniu ryšiu susijusių žodžių *liūdną* ir *žinią*: žodžio *liūdną* nurodyta vyriškoji giminė.

Liūdną jis mums pranešė žinią: šuo nebegrįžo į namus.

TEXT	LEMMA	POS	TAG
Liūdną	liūdnas	ADJ	bdv.nelygin.vyr.vns.G.
jis	jis	PRON	įv.vyr.vns.V.
mums	mums	PRON	įv.dgs.N.
pranešė	pranešti	VERB	vksm.asm.tiesiog.būt-k.vns.3.
žinią	žinia	NOUN	dkt.mot.vns.G.
:	:	PUNCT	skyr.
šuo	šuo	NOUN	dkt.vyr.vns.V.
nebegrįžo	nebegrįžo	VERB	vksm.asm.tiesiog.būt-k.vns.3.
į	į	ADP	prl.G.
namus	namas	NOUN	dkt.vyr.dgs.G.
.	.	PUNCT	skyr.

99 pav. Lietuvių kalbos sakinio analizė, atlikta su *SpaCy 3.0*: didelės apimties modelis

4.3. Skyriaus išvados

Sakinio sintaksinei struktūrai pavaizduoti buvo sukurti du iš principo skirtingi metodai: frazių ir priklausomybių. Pirmasis – frazių – metodas labiau tinka kalboms, turinčioms griežtą žodžių tvarką, nes sakinio struktūra susijusi su linijiniu žodžių išsidėstymu sakinyje. Antrasis – priklausomybių – geriau atspindi tų kalbų savybes, kuriose žodžiai sakinyje išsidėsto laisvai. Pastaruoju metu bandoma jungti abi metodikas – sukurtos universaliosios priklausomybės. Tačiau tokiu atveju atsiranda neprojektyvūs sakiniai, t. y. tokie, kurių sintaksinėje struktūroje matyti susikertančios linijos. Lietuvių kalbos sintaksiškai anotuotame tekstyne ALKSNIS sakiniai vaizduojami universaliųjų priklausomybių principu.

Sintaksiniams, kaip ir morfologiniams, analizatoriams sukurti naudojami du pagrindiniai metodai: taisyklėmis pagrįstas ir statistinis. Formalių gramatikų taisyklės nusako, kaip gauti sakinio sintaksinę struktūrą. Šiuo principu veikiantis lietuvių kalbos sintaksinis analizatorius suformuoja sakinio struktūrą frazių gramatikos metodu ir vėliau ją transformuoja į priklausomybių gramatikos medį. VDU tinklalapyje šis analizatorius buvo prieinamas iki 2020 m. vasario mėn., šiuo metu jokia atnaujinta versija kol kas nepateikta. Kuriant lenkų kalbos sintaksinį analizatorių buvo daugiausia atsižvelgiama į tai, kaip sakinį nagrinėja žmogus.

Statistiniais metodais veikiantis lietuvių kalbos sintaksinis analizatorius pateikiamas internete kaip *UDPipe* lietuvių kalbos modulis, kurio apmokymui buvo naudojamas lietuvių kalbos sintaksiškai anotuotas tekstynas ALKSNIS. Netiksli analizė gaunama dažniausiai tiems sakiniams, kuriuose atsispindi specifiniai lietuvių kalbos bruožai, pvz., anglų kalbai nebūdingas žodžių išsidėstymas ir kt.

5. SKAITMENINĖ GRAMATIKA

Skyriaus antraštėje pavartotas terminas *skaitmeninė gramatika* yra vienas iš tų naujų terminų, kurie yra paplitę ir, galima sakyti, „madingi“, bet jų reikšmė įvairiuose tekstuose labai nevienoda. Todėl šiame skyriuje bus daug vietos skiriama pačiai sąvokai *skaitmeninė gramatika* aiškinti ir nurodomi keli jos pavartojimo atvejai. Taip pat pateikiami iliustraciniai pavyzdžiai apie bandymus sukurti skaitmenines gramatikas daugeliui kalbų, aptariami jų teigiami ir neigiami bruožai bei panaudojimas, analizuojamos problemos, su kuriomis susiduria skaitmeninių gramatikų kūrėjai. Be to, aprašomi ir pirmieji bandymai sukurti lietuvių kalbos skaitmeninę gramatiką.

Terminu *skaitmeninė gramatika* šiuo metu pasaulyje kartais vadinami gana skirtingi dalykai. Kompiuterinėje lingvistikoje tai yra formalus gramatikos taisyklių aprašas, kuris turbūt tiksliausiai atspindi esmę. Kiti pavartojimo atvejai: elektronine forma išleisti gramatikos vadovėliai ir kompiuterizuotos apžvalginės gramatikos.

5.1. Plačiajai visuomenei skirti kitų kalbų elektroniniai gramatikos leidiniai

Šiais laikais *skaitmeninės gramatikos* pavadinimu kartais leidžiami elektroniniai gramatikos vadovėliai, skirti tiek užsienio kalbai (pvz., vokiečių), tiek gimtajai (pvz., švedų) mokytis.

Leidykla CHRISTOS KARABATOS (59 interneto nuoroda[98]) išleido kompaktinę plokštelę graikams, besimokantiems vokiečių kalbos, *Meine Grammatik – Digital* (60 interneto nuoroda[99]).

Buvo dar vienas skaitmeninės gramatikos, skirtos užsieniečiams, besimokantiems vokiečių kalbos, pavyzdys – interneto puslapis *Deutsch-Digital*, skirtas olandams (61 interneto nuoroda[100]). Veiksmažodžių asmenavimo pavyzdys pateiktas 100 pav.

[98] Prieiga internete: http://www.karabatos.gr/ [žiūrėta 2022-11-22].

[99] Prieiga internete: http://www.karabatos.gr/de/meine-grammatik-digital-cd-rom-f%C3%BCr-interaktive-whiteboards [žiūrėta 2022-11-22].

[100] Prieiga internete: http://deutsch-digital.nl/index_grammatica.htm [žiūrėta 2014-05-30].

Deutsch-Digital

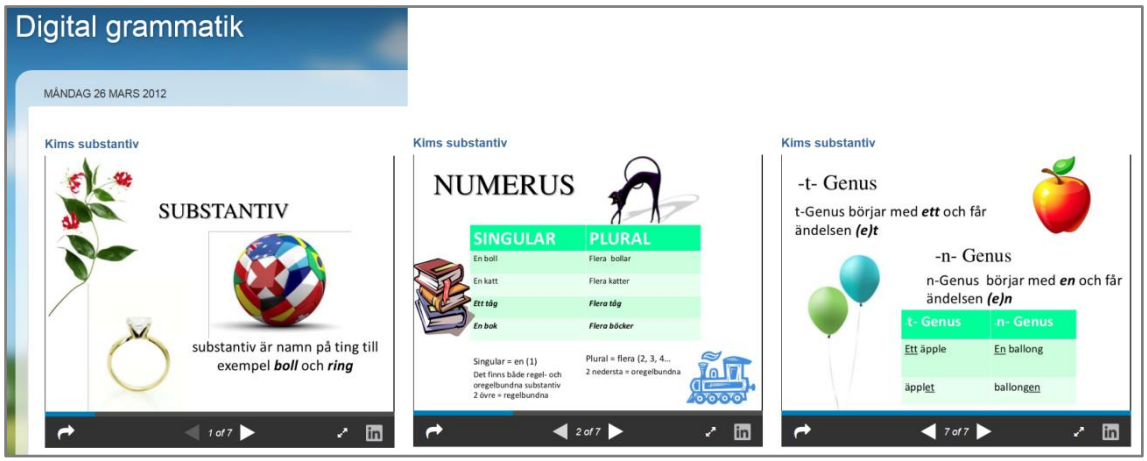
2 werkwoord met stam op –d of –t

Wanneer de stam van een werkwoord eindigt op –d of –t en bovendien bij de werkwoorden atmen, rechnen, regnen, zeichnen, öffnen, leugnen e.a. treedt de regel van de extra –e in werking: elke uitgang moet dan met een –e beginnen. Dus

o.t.t.	arbeiten	rechnen	bieten
	arbeit e	rechn e	biet e
	e st	e st	e st
	e t	e t	e t
	en	en	en
	e t	e t	e t
	en	en	en
	en	en	en
Gebiedende wijs	*arbeit **e** !	*rechn **e** !	*biet **e** !
	arbeit e t !	rechn e t !	biet e t !
	arbeit en Sie !	rechn en Sie !	biet en Sie !
Voltooid deelw.	ge arbeit e t	ge rechn e t	geboten

100 pav. Olandams skirtos skaitmeninės vokiečių kalbos gramatikos pavyzdys (61 interneto nuoroda)

Švedijoje buvo paruoštos skaidrės, pavadintos *Skaitmenine gramatika*, padedančios mokytis gimtosios kalbos (62 interneto nuoroda[101]). Jose labai populiariai išdėstytos gramatikos taisyklės, pateikti paveikslėliais iliustruoti gramatinių kategorijų (skaičiaus, linksnio, giminės ir kt.) aiškinimai. 101 pav. parodyta keletas tokių skaidrių pavyzdžių.



101 pav. Švedų kalbos skaitmeninės gramatikos skaidrių pavyzdžiai (62 interneto nuoroda)

[101] Prieiga internete: http://digitalgrammatik.blogspot.com/ [žiūrėta 2022-11-22].

5.2. Apžvalginės gramatikos

Galima išskirti du apžvalginių gramatikų (angl. *Sketch grammar*) rašymo tipus: vienas – analogiškas spausdintoms popieriuje knygoms, kitas – artimas kompiuterinei gramatikai.

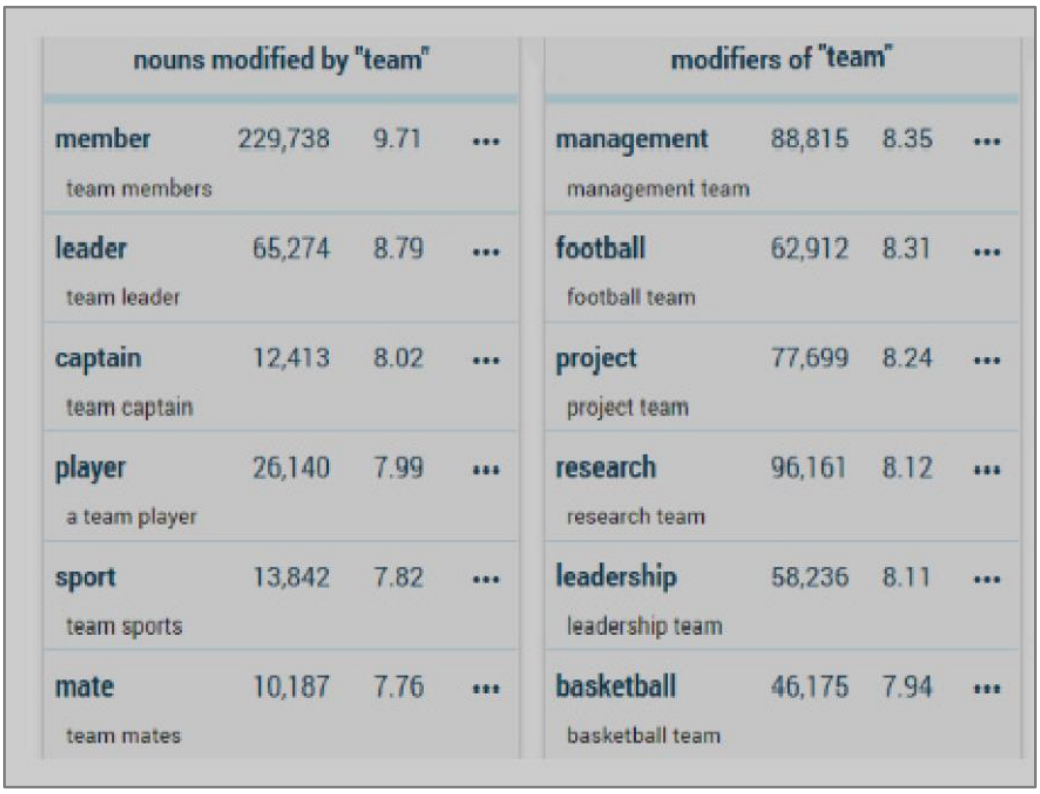
5.2.1. Apžvalginių gramatikų rašymo metodai

Apžvalginė gramatika, analogiška spausdintoms popieriuje knygoms, rašoma tautų kalbomis, tik nuo akademinės gramatikos ji skiriasi keliais bruožais. Pirmiausia, ji yra daug mažesnės apimties, be to, jai, kaip ir žodynams, turi būti nustatytos tikslinės grupės, kurioms ji skiriama. Apžvalginėje gramatikoje pateikiamos pagrindinės žinios apie kalbą, parodomas tos kalbos gramatikos išskirtinumas (Donohue 2011: 1). Anglų kalbos apžvalginė gramatika parašyta kaip pavyzdys ir skirta tikslinei grupei, kurią sudaro kalbininkai, studijuojantys mažai ištyrinėtas pasaulio kalbas. Labiausiai ji tinka studentams, kurie per vieno ar dviejų semestrų kursą planuoja parengti deskriptyvinę morfosintaksę tos kalbos, kuri nėra jų gimtoji. Anglų kalbos apžvalginė gramatika parašyta taip, tarsi ši kalba būtų mažai tyrinėta. Pagrindinis jos tikslas – pasiūlyti kalbininkams minčių ar idėjų, kaip kalbos gali būti aprašomos išsamumo, apimties ir pateikimo aspektais (Payne 2010).

Kitas apžvalginių gramatikų rašymo metodas yra formalus taisyklių aprašas, naudojamas kompiuterinio kalbos apdorojimo metu. Tačiau pats informacijos pateikimo principas išlieka – tai yra supaprastinta gramatika. Ji skirta tekstynams analizuoti ir žodžių apžvalgoms (angl. *word sketch*) sudaryti, kurios nuo kolokacijų skiriasi tuo, kad į vieną puslapį sudedama žodžio funkcionavimo kalboje suvestinė, gauta automatiškai apdorojant tekstyną. Žodžio apžvalgoje dažniausiai pateikiama trijų rūšių informacija apie jį: pagrindinis žodžių junginio dėmuo (angl. *headword*), sintaksinė funkcija, kolokacija, pvz.: *man*, *modifier*, *young*. Kartais po jų dar nurodoma pagrindinio žodžio bei kolokacijos vieta tekstyne, pvz.: *man*, *modifier*, *young*, *104*, *103* (63 interneto nuoroda[102]).

[102] Prieiga internete: https://en.wikipedia.org/wiki/Word_sketch [žiūrėta 2022-11-22].

Žodžių apžvalgas galima sudaryti naudojantis programine įranga *Sketch Engine* (64 interneto nuoroda[103]). 102 pav. parodytas anglų kalbos žodžio *team* apžvalgos fragmentas, išsamiau ji pateikiama 8 priede.



nouns modified by "team"				modifiers of "team"			
member team members	229,738	9.71	•••	management management team	88,815	8.35	•••
leader team leader	65,274	8.79	•••	football football team	62,912	8.31	•••
captain team captain	12,413	8.02	•••	project project team	77,699	8.24	•••
player a team player	26,140	7.99	•••	research research team	96,161	8.12	•••
sport team sports	13,842	7.82	•••	leadership leadership team	58,236	8.11	•••
mate team mates	10,187	7.76	•••	basketball basketball team	46,175	7.94	•••

102 pav. Žodžio *team* apžvalgos, paruoštos programine įranga *Sketch Engine*, fragmentas (65 interneto nuoroda[104])

Kad žodžius būtų galima suskirstyti pagal sintaksines funkcijas, pvz.: papildinys, pažyminys ir kt., turi būti parašyta apžvalginė gramatika, t. y. taisyklių, pagal kurias automatiškai nustatomi žodžių ryšiai, rinkinys. 103 pav. pateiktas tokios taisyklės pavyzdys (66 interneto nuoroda[105]).

```
1:"VB.?" [tag="DT|PRP$"]{0,1} "JJ.?"{0,3} "NN.*"{0,2} 2:"NN.*"
```

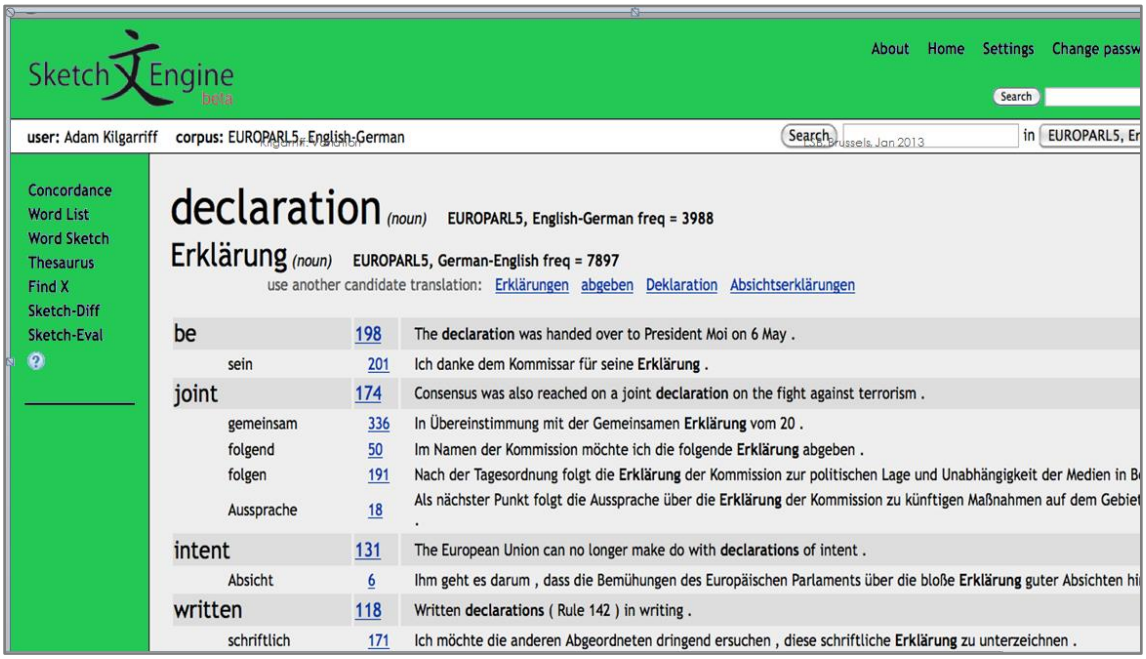
103 pav. Apžvalginės gramatikos taisyklė (66 interneto nuoroda)

[103] Prieiga internete: https://www.sketchengine.eu/what-can-sketch-engine-do/ [žiūrėta 2022-11-22].

[104] Prieiga internete: https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/ [žiūrėta 2022-11-22].

[105] Prieiga internete: https://www.sketchengine.eu/documentation/writing-sketch-grammar/ [žiūrėta 2022-11-22].

Ši taisyklė apima atvejus, kai analizuojamas žodis (žymimas kaip *1:*) yra veiksmažodis *"VB.?"* ir po jo gali eiti apibrėžiklis (angl. *determiner*) arba prielinksnis *[tag="DT|PRP$"]*, pasikartojantys vieną ar nulį kartų *{0,1}*. Po jų gali būti iki trijų būdvardžių *"JJ.?" {0,3}* (būdvardis anotuojant anglų kalbos tekstyną dažnai žymimas *JJ*, pvz., *Penn-treebank* tekstyne, 19 pav.) ir iki dviejų daiktavardžių *"NN.*" {0,2}*. Paskutinė taisyklės dalis *2:"NN.*"* rodo, kad ieškoma analizuojamo žodžio *1: "VB.?"* ryšio su daiktavardžiu. Taip parengtos gramatikos nėra labai tikslios: gali būti ir neatpažintų ryšių, kai žodžiai susiję, ir neteisingai nustatytų ryšių tarp žodžių, kurie iš tikrųjų jo neturi (66 interneto nuoroda). Netikslaus informacijos pateikimo pavyzdys parodytas 104 pav., kurį aprašo pats *Sketch Engine* įkūrėjas (Kilgarriff 2013: 22). Numeriu 198 pažymėtoje eilutėje anglų kalbos žodis *be* (konkrečiai šiame sakinyje *was*) iš tikrųjų susijęs su žodžiu *declaration* (*declaration was handed*). Kaip jo (žodžio *be*) atitikmuo vokiečių kalboje pateikiamas žodis *sein*, kurio viena reikšmių yra *būti*. Tačiau numeriu 201 pažymėtoje eilutėje pateiktas junginys *seine Erklärung*, kur žodis *sein* turi kitą reikšmę: jis pavartotas kaip įvardžio *jo* (vok. *sein*) moteriškoji giminė *seine*. Numeriais 118 ir 171 pažymėtose eilutėse pateikta visiškai tiksli informacija: anglų kalbos žodžių junginio *written declaration* vokiškas atitikmuo iš tikrųjų yra *schriftliche Erklärung*.



104 pav. *Europarl5* anglų ir vokiečių kalbų žodžių *declaration* ir *Erklärung* apžvalgos pavyzdys (parengta pagal Kilgarriff 2013: 22)

5.2.2. Tekstyno modelių analizė

Apžvalginė lietuvių kalbos gramatika buvo parašyta kaip pagalbinė priemonė kuriant *Internetinį besimokančiojo lietuvių kalbos žodyną* ir skirta tekstyno modelių analizei automatizuoti. Žodynui buvo sudaromi žodžių modeliai, naudojantis indukcine[106] (67 interneto nuoroda[107]) tekstyno modelių analize (angl. *corpus pattern analysis* – CPA). Čia vertėtų kiek plačiau paaiškinti, kas yra CPA.

Pagrindinė XXI a. leksikografų užduotis – pateikti dabartinę, faktinę tam tikros kalbos žodžių vartoseną. Žodynuose nepasakyta, kaip nustatyti skirtingas to paties žodžio reikšmes, pvz.: *these patients are treated with antibiotics* ir *as sisters and daughters women are treated with respect*, pirmajame sakinyje žodis *treat* reiškia *gydyti*, o antrajame sakinyje tas pats žodis *treat* reiškia *elgtis*. Žmonės tai žino, tačiau užsieniečiams ar kompiuterių programoms tokiais atvejais gali iškilti problemų. Todėl buvo sukurti naujo tipo žodynai, kuriuose dėmesys sutelktas labiau į vartojimą negu į reikšmę. Jiems parengti buvo naudojama tekstyno modelių analizė. Į klausimą *Kaip nustatyti žodžio reikšmę daugiareikšmiškumo atveju?* dažniausiai atsakoma: *Iš konteksto*. Taigi, daugiareikšmiam anglų kalbos žodžiui *toast*[108] yra geriau sudaryti, viena vertus, tokius leksikos rinkinius kaip *duona*, *batonas*, *bandelė* ir, kita vertus, tokius kaip *žmonės*, *jų sėkmės ir pergalės*, *jų sveikata ir ateitis*, *jei jie gyvi*, *ir jų atminimas*, *jei mirę*, nei suteikti semantinį požymį *maistas*, nes ne visas maistas yra skrudinamas (Hanks 2004: 90–92). Analizuojant tekstyną, sudaromi modeliai, atspindintys žodžių vartojimą ir atskiroms kalbos dalims jie yra nevienodi. Pavyzdžiui, daiktavardžių modeliai sudaromi grupuojant dažnai vartojamus posakius, o veiksmažodžių modeliai apima ne tik jų valentingumą, bet ir subvalentinius požymius, pvz., anglų kalbos apibrėžiklį: *take place* ir *take his place* priklausys skirtingiems modeliams (Hanks 2004: 87).

Grįžtant prie lietuvių kalbos modelių reikia pasakyti, kad *Internetinis besimokančiojo lietuvių kalbos žodynas* – tai leksinė duomenų bazė, skirta užsieniečiams, besimokantiems lietuvių kalbos. Jis buvo rengiamas tekstyno modelių analizės pagrindu. Pasinaudojant nedideliu *Besimokančiųjų tekstynu*, apimančiu beveik

[106] Indukcija – išvadų gavimas iš atskirų teiginių.

[107] Prieiga internete: https://www.zodynas.lt/terminu-zodynas/I/indukcija [žiūrėta 2022-11-22].

[108] 1) cook food by exposure to a grill or fire; 2) raise one's glass and drink in honour of someone or something (Hanks 2004: 91).

700 000 žodžių, kurį sudaro elektroninių tekstų, skirtų A1, A2, B1, B2 lygiams, rinkiniai (68 interneto nuoroda[109]), buvo sudaromi žodžių modeliai. Autoriai nurodo, kad visi tekstai *Besimokančiųjų tekstyne* anotuoti morfologiškai automatiniu būdu, todėl gali pasitaikyti tam tikrų netikslumų, tačiau jų skaičius yra nedidelis ir nepadaro įtakos bendroms tendencijoms (Boizou, Kovalevskaitė, Rimkutė 2020: 233–236). Modeliai buvo sudaryti tik tiems žodžiams, kurie tekstyne pasitaikė ne mažiau kaip 100 kartų. Žodynas apima 3 000 leksinių vienetų. Antraštinių žodžių sąrašą sudaro daiktavardžiai, veiksmažodžiai (išskyrus pagalbinius ir modalinius), būdvardžiai, prieveiksmiai (išskyrus parodomuosius: *čia*, *ten*, *dabar* ir kt.) ir kai kurie skaitvardžiai (*šimtas*, *tūkstantis*, *milijonas*). Kadangi žodžio reikšmė susijusi su specifine jo leksine ir gramatine aplinka, tekstyno analizės metodas buvo šiek tiek modifikuotas, kad būtų galima gauti prasminius modelius. Tam tikslui buvo sukurta *Lietuvių kalbos apžvalginė gramatika – Lithuanian Sketch Grammar* (Kovalevskaitė ir kt. 2020: 247).

5.2.3. *Lietuvių kalbos apžvalginė gramatika*

Morfologiškai anotavus *Besimokančiųjų tekstyną* programine įranga *semantika.lt*, buvo galima parengti morfologija pagrįstą apžvalginę lietuvių kalbos gramatiką. Taisyklės buvo sudaromos remiantis tikėtinais tipiškais priklausomais žodžiais kiekvienai atskirai kalbos daliai:

a) VEIKSMAŽODŽIAMS – įvairių linksnių (išskyrus šauksmininką) daiktavardžiai / įvardžiai, būdvardis (veiksmažodžiui *būti*), prielinksnis, bendratis, jungtukas;
b) DAIKTAVARDŽIAMS – prepoziciniai būdvardžiai / dalyviai, suderinti linksniu, prepoziciniai kilmininko linksnio daiktavardžiai, kai kurie postpoziciniai priklausomi naudininko linksnio žodžiai, pvz., *įtaka kam*, arba kilmininko linksnio žodžiai, susieti per prielinksnį, pvz., *priemonė nuo ko*;
c) BŪDVARDŽIAMS – prieš juos esantys prieveiksmiai, kai kurie po jų esantys priklausomi įnagininko linksnio žodžiai, pvz., *įdomus kuo*; bendratis, pvz., *svarbu žinoti*; prielinksnis, pvz., *greitesnis už ką* ir kt.;
d) PRIEVEIKSMIAMS – prieš juos esantys prieveiksmiai, pvz., *labai akivaizdžiai* ir pan.

[109] Prieiga internete: https://baltnexus.lt/lt/baltistikos-projektas, skirtukas *3.2.2. Interaktyvios mokymo priemonės* [žiūrėta 2022-11-22].

Lietuvių kalbos apžvalginei gramatikai buvo sudaryta 14 taisyklių, aprašančių dvejopus ryšius tarp žodžių, pvz., *turi daiktavardinį pažyminį / yra daiktavardinis kieno pažyminys* ir pan. Visas sąrašas pateikiamas 9 priede. Nustatant ryšius, buvo įvertinama, kad tarp tiesioginiu ryšiu susijusių žodžių gali įsiterpti kiti žodžiai, tačiau tų įsiterpusių žodžių kiekis yra ribotas. Didžiausias leidžiamas atstumas buvo trys žodžiai, ieškant veiksmažodinių junginių, kitais atvejais – vienas ar net nė vieno žodžio. Dar viena ypatybė, būdinga veiksmažodiniams junginiams, yra ta, kad priklausomi žodžiai gali būti tiek prieš veiksmažodį, tiek po jo. Kitų kalbos dalių junginiams nustatyta fiksuota pozicija prieš arba po pagrindinio žodžio (Kovalevskaitė ir kt. 2020: 248).

Besimokančiųjų tekstynas nėra sintaksiškai anotuotas, todėl, remiantis apžvalgine gramatika, gali būti nustatyti ryšiai tarp žodžių, kurie sintaksiškai nėra susiję. Be to, kai kurie linksniai yra daugiareikšmiai, ypač kilmininkas, kuris daugiausia naudojamas kaip pažyminys, tačiau kartais gali būti ir papildinys. Žodžių apžvalgai buvo atrenkami tik tie junginiai, kurie aprašyti apžvalginės gramatikos taisyklėmis. (Iš esmės apžvalginė gramatika yra tarsi filtras, paruošiantis duomenis leksikografams, rengiantiems *Internetinį besimokančiojo lietuvių kalbos žodyną*.) Toliau pagal atrinktus žodžių junginius sudaromi tekstyno modeliai, atspindintys gramatinę (sintaksines funkcijas bei linksnių ir veiksmažodžio formų morfologines kategorijas), leksinę (žodžius, kolokacijas) ir semantinę informaciją. Požymis, t. y. kolokacija, gramatinė forma ar sintaksinė funkcija, turi pasitaikyti tekstyne ne mažiau kaip tris kartus, kad būtų įtrauktas į modelį. Dažnai pasitaikantys modeliai buvo surasti pasitelkus apžvalginę gramatiką.

Kartais užtenka vien valentingumo, kad būtų nustatyti reikšmės skirtumai, pvz., veiksmažodis *reikšti* turi dvi prasmes: *nurodyti*, *žymėti* ir *turėti vertę*. Pirmuoju atveju modelis atrodytų kaip pateikta 105 pav., antruoju – kaip 106 pav. (Kovalevskaitė ir kt. 2020: 249).

[Sub_nom] [REIKŠTI] [Obj_acc]: Geltona spalva reiškia saulę.

105 pav. Žodžio *reikšti* modelis, kai jo prasmė – *nurodyti*, *žymėti* (parengta pagal Kovalevskaitė ir kt. 2020: 249)

[Sub_nom] [Obj_dat] [REIKŠTI] [Obj_acc]: Ką Tau reiškia pokalbis?

106 pav. Žodžio *reikšti* modelis, kai jo prasmė – *turėti vertę*
(parengta pagal Kovalevskaitė ir kt. 2020: 249)

Mokantis lietuvių kalbos, kuriai būdingas didelis kaitomumas, svarbu suprasti linksnių ir kitų gramatinių formų reikšmę, todėl sudaromas daugiasluoksnis modelio aprašas, t. y. gramatiniai, semantiniai ir leksiniai duomenys pateikiami atskirai. 107 pav. parodytas žodžio *arbata* aprašas (Kovalevskaitė ir kt. 2020: 250).

"gramForm": **[ARBATA] [su AtrN_ins]**
"semForm": **[Mod] [maistas]**
"collocates": **[ARBATA] [su citrina]**

107 pav. Žodžio *arbata* aprašas (parengta pagal Kovalevskaitė ir kt. 2020: 250)

Apžvalginės gramatikos pagrindu sukurtas *Mokomasis lietuvių kalbos vartosenos leksikonas* prieinamas internete VDU portale *Kalbu* (69 interneto nuoroda[110]). Tai visiškai naujas leksikografinis resursas, nes jo antraštynas neperimtas iš jokio žodyno, o parengtas remiantis konkrečiu tekstynu (Kovalevskaitė, Rimkutė 2022: 155). Visą leksikono antraštyną sudaro apie 3 700 vienetų. Kitų tekstynais paremtų mokomųjų žodynų rengėjai antraštynui formuoti naudojo didesnį tekstyną, pvz., estų mokomajam žodynui naudotas bendrasis 250 mln. žodžių tekstynas *Estonian Reference Corpus*, iš jo gautas 5 000 leksinių vienetų antraštynas (Kovalevskaitė, Rimkutė 2022: 176–177).

5.3. Formalus gramatikos taisyklių aprašas

Skaitmeninės gramatikos, skirtos kalbai kompiuterizuoti, buvo pradėtos kurti, kai paaiškėjo, kokių sunkumų iškyla apdorojant kalbas statistiniais metodais: dėl didelės duomenų apimties būna sunku numatyti proceso baigtį ir beveik neįmanoma jo kontroliuoti. Naudojant statistinius metodus, greitai sukuriamos galingos sistemos, gebančios apdoroti nepaprastai dideles žodžių apimtis, tačiau tik su viena sąlyga – turi

[110] Prieiga internete: https://kalbu.vdu.lt/ [žiūrėta 2022-11-22].

būti toleruojamas tam tikras kiekis klaidų. Skaitmeninės gramatikos – tai kruopščiai sudaromi ir patikrinti kalbos ištekliai, kurių pateikiama informacija yra itin aukšto tikslumo (70 interneto nuoroda[111]).

5.3.1. *Gramatinė struktūra*

Skaitmeninėms gramatikoms rašyti buvo sukurta speciali programinė įranga *Gramatinė struktūra* (angl. *Grammatical Framework*). Ją parengė Geteborgo universiteto (University of Gothenburg) profesorius Aarne Ranta ir pirmą kartą ji buvo įdiegta 1998 m. Grenoblyje, *Xerox* tyrimų centre. Naudojantis ja galima atlikti tiek teksto gramatinę analizę, tiek jo sintezę iš karto keliomis kalbomis, nes remiamasi nuo konkrečios kalbos nepriklausomu reikšmės atvaizdavimu[112] (71 interneto nuoroda[113]). *Gramatinės struktūros* paskirtis – formalizuoti pasaulio kalbų gramatikas, kad jas būtų galima panaudoti apdorojant kalbas kompiuteriu. Tai yra atvirojo kodo programinė įranga, kuria visi gali naudotis nemokamai (Ranta 2015). Šiuo metu, remiantis *Gramatine struktūra*, kuriamos beveik 40 kalbų skaitmeninės gramatikos, tačiau išsamiausi duomenys yra surinkti apie anglų kalbą. Reikia atkreipti dėmesį, kad ir pats aprašas atspindi anglų kalbos ypatybes, nors stengiamasi atsižvelgti ir į kitų kalbų specifinius bruožus, pvz., didelį kaitomumą. Išsamiau šis klausimas bus aprašytas 5.3.2.2 poskyryje.

Viena svarbiausių *Gramatinės struktūros* ypatybių yra ta, kad sintaksė čia suskaidyta į konkrečią ir abstrakčią. Abstrakčioje sintaksėje sukaupti nuo kalbos nepriklausomi duomenys, ji remiasi semantika, o konkreti sintaksė aprašo kiekvienai atskirai kalbai būdingą sintaksinę ir leksinę abstrakčios sintaksės realizaciją (Grūzītis, Dannélls 2017: 6).

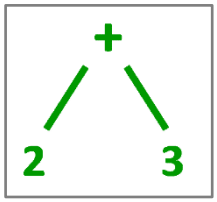
Abstrakti sintaksė – tai priklausomybių gramatikos medžių rinkinys (Ranta 2009: 5). Sakinys pateikiamas sąvokų struktūros medžiu, kuris tarnauja kaip apibendrintas semantinis (reikšmės) formatas. Šis pavaizdavimas yra bendras visoms kalboms ir apibrėžia sakinio prasmę. Konkreti sintaksė pateikia abstrakčios sintaksės reikšmę kiekvienai atskirai kalbai būdingu pavidalu. Sistemos universalumą rodo tai,

[111] Prieiga internete: https://www.digitalgrammars.com/technology [žiūrėta 2022-11-22].

[112] "Grammatical Framework (GF) is a programming language for writing grammars of natural languages. GF is capable of parsing and generating texts in several languages simultaneously while working from a language-independent representation of meaning" (71 interneto nuoroda).

[113] Prieiga internete: https://en.wikipedia.org/wiki/Grammatical_Framework [žiūrėta 2022-11-22].

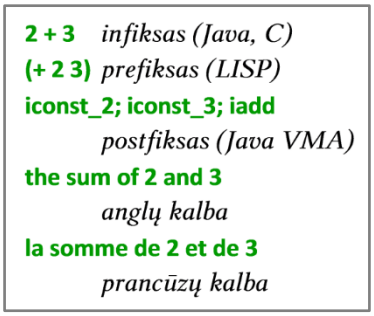
kad kalbos gali būti tiek tautų, tiek formalios (Hallgren ir kt. 2015: 42), pvz., programavimo. Abstrakčiai ir konkrečiai sintaksei iliustruoti pateikiamas matematinės operacijos pavyzdys: *Sudėti du ir tris*. Priklausomybių gramatikos medžio viršūnėje įrašomas tarinys. Abstrakčioje sintaksėje tai bus tiesiog ženklas + (108 pav.). Tarinį išplečiantys žodžiai taip pat vaizduojami simboliais – skaičiais. Taigi, šis abstrakčios sintaksės medis apima tris sąvokas – sudėties operaciją ir du skaičius: du ir trys (Ranta 2011: 29).



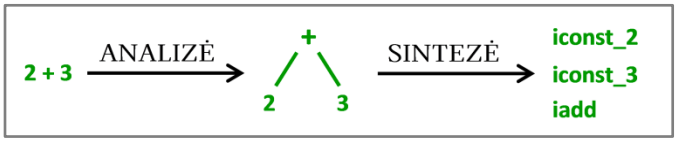
108 pav. Abstrakti sintaksė: sąvokų medis sakiniui *Sudėti du ir tris* (parengta pagal Ranta 2011: 29)

109 pav. pateikiamas konkrečios sintaksės pavyzdys: keliomis skirtingomis kalbomis, todėl ir skirtingu pavidalu, parašyta tai, kas yra užkoduota abstrakčios sintaksės medyje (108 pav.). 109 pav. aiškiai matyti, kad kalbos gali būti labai įvairios, t. y. tiek programavimo, tiek tautų kalbos. *Java* ir *C* programavimo kalbose sudėties aritmetinė operacija užrašoma pliuso ženklu ir nurodoma kaip infiksas, t. y. tarp skaičių arba kintamųjų (kurie gali būti užrašyti raidžių ir / ar skaičių rinkiniu). Programavimo kalboje *LISP* aritmetinė operacija parašoma prefiksu, t. y. prieš skaičius ar kintamuosius, su kuriais ji turi būti atlikta. *Java VMA* kalboje operacija nukeliama į postfiksą, t. y. ji užrašoma po skaičių ar kintamųjų. Pats užrašas šioje programavimo kalboje pateikiamas labiau įprasta žmogui forma, t. y. žodžiais, tiksliau, jų sutrumpinimais: *iadd* reiškia *integer add*, *iconst_2* – *integer constant 2* ir t. t. Objektai, su kuriais turi būti atliekama operacija, ir pati operacija atskiriami kabliataškiais. Abstraktaus medžio reikšmė (108 pav.) gali būti perteikiama ir tautų kalbomis: anglų, prancūzų ir kitomis. Visi šie užrašai labai skiriasi, bet jie turi tą pačią prasmę – apima tris sąvokas: sudėties operaciją ir du skaičius. Visos šios sąvokos atsispindi kiekvienoje iš konkrečių kalbų. Ir joms visoms vienodai gerai tinka skaitmeninės gramatikos aprašas (Ranta 2011: 29).

Ryšiai tarp sintaksės konkrečios ir abstrakčios dalių yra dvikrypčiai. Iš bet kurios konkrečios kalbos galima gauti bendrą semantinę sakinio struktūrą, t. y. atlikti sakinio analizę; ir iš abstraktaus pavidalo (medžio) galima sugeneruoti bet kurios kalbos sakinį, turintį tą pačią prasmę kaip ir abstrakčios sintaksės medis (110 pav.).

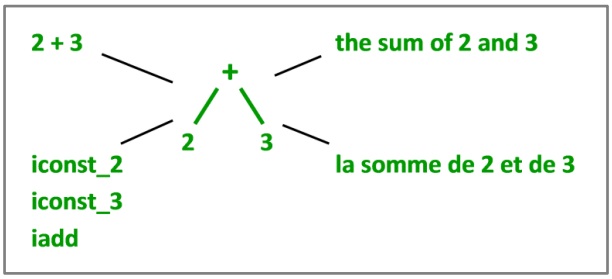


109 pav. Konkreti sintaksė: sąvokų medžio sakiniui *Sudėti du ir tris* realizacija įvairiomis kalbomis (parengta pagal Ranta 2011: 29)



110 pav. Dvikrypčiai ryšiai tarp sintaksės konkrečios ir abstrakčios dalių (parengta pagal Ranta 2011: 30)

Pagrindinis *Gramatinės struktūros* bruožas yra daugiakalbiškumas. Tai reiškia, kad vieną abstraktų sakinio aprašą atitinka daug skirtingų konkrečių to paties sakinio pavidalų (111 pav.).



111 pav. Daugiakalbiškumas: vienas abstraktus pavaizdavimas ir daug konkrečių atitikmenų (parengta pagal Ranta 2011: 33)

5.3.2. *Gramateka* – skaitmeninių gramatikų biblioteka

Naudojant *Gramatinę struktūrą* sukurta daug skaitmeninių gramatikų ir jos visos kaupiamos gramatikos išteklių bibliotekoje – ***Gramatekoje*** (angl. *Resource Grammar Library* – RGL). 2022 m. duomenimis, ji apėmė 38 kalbas, tarp jų ir estų bei latvių, tačiau lietuvių kalbos šiame sąraše dar nėra (72 interneto nuoroda[114]). 2001 m. *Gramateką* tesudarė vos trys kalbos: anglų, švedų ir rusų. 2011 m. joje jau buvo 20-ies kalbų skaitmeninės gramatikos (Ranta 2011: 48).

[114] Prieiga internete: http://www.grammaticalframework.org/lib/doc/synopsis/ [žiūrėta 2022-11-22].

A. Ranta (2013) populiariai aprašė pačią idėją. Visą *Gramateką* galima įsivaizduoti kaip sudarytą iš dviejų didelių sričių: žodžių ir sintaksės. Kiekviena šių sričių turi po du skyrius: bendrąjį ir specifinį. Remiamasi tuo, kad įvairių kalbų gramatikose didelė dalis yra bendra, nepaisant pastebimų skirtumų. Pavyzdžiui, visos šiuo metu *Gramatekoje* esančios kalbos turi daiktavardžio kategoriją, bet apibrėžimas, kas tiksliai yra daiktavardis, įvairiose kalbose skiriasi: anglų kalbos daiktavardis turi keturias formas (vienaskaitos ir daugiskaitos vardininką ir kilmininką); prancūzų kalboje yra tik dvi formos (vienaskaita ir daugiskaita), bet čia daiktavardžiai turi giminę, kurios nėra anglų kalboje; vokiečių kalbos daiktavardžių yra aštuonios formos ir trys giminės; kinų kalba teturi vieną vienintelę daiktavardžio formą, o suomių kalboje jų net dvidešimt šešios ir t. t. Išsamiau šis klausimas aptariamas 5.3.2.2 poskyryje. Visų *Gramatekoje* esančių kalbų aprašas pateikiamas anglų kalba.

5.3.2.1. Bendrasis žodžių srities skyrius

Bendrajame žodžių srities skyriuje aprašomi visi žodžiai nurodant jų charakteristikas. Ši *Gramatekos* dalis yra vienoda visoms kalboms. Visos joje esančios kalbos turi bendrą žodžių klasifikaciją, skiriasi tik jų morfologija. Iš pradžių žodžiai skaidomi į savarankiškus (angl. *content words*) ir nesavarankiškus (angl. *structural words*). 112 pav. pateikta lentelė, aprašanti savarankiškus žodžius.

GF name	text name	example	inflectional features	inherent features
N	noun	*house*	number, case	gender, classifier
PN	proper name	*Paris*	case	gender
A	adjective	*blue*	gender, number, case, degree	position
V	verb	*sleep*	number, person, tense, aspect, mood	subject case
Adv	adverb	*here*	(none)	adverb type (place, time, manner)
AdA	adadjective	*very*	(none)	(none)

112 pav. Savarankiškų žodžių charakteristikos (Ranta 2013)

Pirmajame stulpelyje nurodomas sutrumpintas kategorijos pavadinimas *GF name*[115], kuris naudojamas programinės įrangos kode. Antrajame stulpelyje parašyta ta pati kategorija, kaip ji atrodo žmonėms skirtame tekste. Trečiajame stulpelyje

[115] GF – Grammatical Framework (Gramatinė struktūra).

pateikiami žodžių pavyzdžiai. Ketvirtajame – surašyti kaitybiniai žodžio požymiai, o penktasis – apima nekaitomus jo požymius. Iš lentelės matyti, pvz., kad daiktavardžiai kaitomi skaičiais ir linksniais, o giminė yra pastovus požymis ir pan. Šioje lentelėje yra apibendrinti visų *Gramatekoje* esančių kalbų duomenys ir tokia lentelė užpildoma kiekvienai kalbai. Galima teigti, kad jei kalba turi giminės kategoriją, tai daiktavardžiai yra nekaitomi giminėmis, t. y. giminė yra pastovus daiktavardžio požymis, o būdvardžiai kaitomi giminėmis. Ir tai lemia ne atskiros kalbos savybės, o sintaksė: būdvardžiai pažymi daiktavardžius ir turi pažymėti juos visus, nepriklausomai nuo giminės, todėl turi įgyti visas giminės formas (Ranta 2009: 8).

Valentingumas *Gramatekoje* įvertinamas vartojant *pozicijų* sąvoką: nurodomi dvipozicinai, tripoziciniai ir pan. daiktavardžiai, būdvardžiai, veiksmažodžiai. 113 pav. ir 114 pav. pateikti fragmentai iš lentelių, kuriose aprašomas atitinkamai veiksmažodžių ir daiktavardžių bei būdvardžių valentingumas (Ranta 2013).

GF name	text name	example	inherent complement features
V2	two-place verb	*love* (*someone*)	case or preposition
V3	three-place verb	*give* (*something to someone*)	two cases or prepositions
VV	verb-complement verb	*try* (*to do something*)	infinitive form
VS	sentence-complement verb	*know* (*that something happens*)	sentence mood

113 pav. Veiksmažodžių valentingumo išraiška pozicijomis (Ranta 2013)

GF name	text name	example	inherent complement features
N2	two-place noun	*brother* (*of someone*)	case or preposition
N3	three-place noun	*distance* (*from some place to some place*)	case or preposition
A2	two-place adjective	*similar* (*to something*)	case or preposition

114 pav. Daiktavardžių ir būdvardžių valentingumo išraiška pozicijomis (Ranta 2013)

5.3.2.2. Specifinis žodžių srities skyrius

Specifiniame žodžių srities skyriuje nurodomi kiekvienai atskirai kalbai būdingų morfologinių kategorijų parametrai. 115 pav. pateikta lentelė iš anglų kalbos žodžių specifinio skyriaus, aprašanti, kokias reikšmes gali įgyti kiekviena morfologinė kategorija, kuri būdinga anglų kalbai (Ranta 2013), pvz., linksnio gali būti tik dvi reikšmės: vardininkas ir kilmininkas ir pan.

GF name	text name	values
`Number`	number	singular, plural
`Person`	person	first, second, third
`Case`	case	nominative, genitive
`Degree`	degree	positive, comparative, superlative
`AForm`	adjective form	degrees, adverbial
`VForm`	verb form	infinitive, present, past, past participle, present participle
`VVType`	infinitive form (for a VV)	bare infinitive, *to* infinitive, *ing* form

115 pav. Anglų kalbos morfologinės kategorijos, nurodytos žodžių specifiniame skyriuje (Ranta 2013)

Daugumai *Gramatekos* kalbų kaitybinės formos užima labai didelę aprašo dalį. Kiekviena kaitybos paradigma pateikiama lentelėje. Anglų kalbos daiktavardžių linksniavimo paradigma (Ranta 2013) parodyta 116 pav.

form	singular	plural
nominative	*dog*	*dogs*
genitive	*dog's*	*dogs'*

116 pav. Anglų kalbos daiktavardžių kaitybos paradigma (Ranta 2013)

Programinė įranga pagal paradigmas sudaro visas galimas pateikto žodžio formas. Tam naudojama programavimo kalbos funkcija, kuri užrašoma specialiu pavidalu. Pvz., **mkV : Str → V** reiškia funkciją *mkV* (*make Verb* – sudaryti veiksmažodį), t. y. iš eilutės *Str* (*string*) padaryti veiksmažodį *V* (*Verb*). Į eilutės (*Str*) vietą įrašius konkretų anglų kalbos žodį, pvz., *walk*, t. y. **mkV : walk → V** bus gaunama visa paradigma: *walk, walks, walked, walking*. Kita funkcija **mkCl : NP → V → Cl** reiškia sudaryti sakinį (*mkCl – make Clause*), t. y. iš daiktavardinės frazės *NP* (*Noun Phrase*) ir veiksmažodžio *V* (*Verb*) padaryti sakinį *Cl* (*Clause*). Ši funkcija rūpinasi ir galūnių suderinimu: *she walks* ar *they walk* (Ranta 2009, 2).

5.4. Daugiakalbiškumo problemos

Bet kokia programinė įranga, vienu metu apdorojanti daugelio tautų kalbas, turi būti iš karto kuriama kaip labai universali sistema, apimanti visų planuojamų įtraukti kalbų ypatybes, nes „skirtumai tarp įvairių kalbų gali būti labai dideli ir esminiai“[116] (Dąbrowska 2015: 1). „Gramatinės kategorijos sukuriamos atskirai

[116] “Languages differ from each other in profound ways” (Dąbrowska 2015).

konkrečiai kalbai tada, kai jų tai kalbai prireikia, ir visiškai nesirūpinama, ar jos tinka kitų kalbų gramatikoms“[117] (Newmeyer 2008: 1 iš Dąbrowska 2015: 2). Atliekant šiuolaikinius lingvistinius tyrimus ir atidžiau panagrinėjus neseniai aptiktas naujas kalbas, galima pastebėti, kad beveik kiekvienoje jų atsiskleidžia netikėti nauji požymiai[118] (Evans, Levinson 2009: 432).

5.4.1. Universalios gramatikos idėja

Pagrindinė (nors ir ne vienintelė) skaitmeninių gramatikų pritaikymo sritis yra automatinis vertimas. Kuriant daugiakalbes automatinio vertimo sistemas būtų labai patogu turėti gramatiką, kuri apimtų visų pasaulio tautų kalbas. Universalios gramatikos idėjų galima įžvelgti dar XIII a. išsakytame Rodžerio Beikono (Roger Bacon) teiginyje, kad visos kalbos remiasi bendra gramatika[119] (Nordquist 2018). XVII a. Port Rojalio (Port Royal) gramatikos mokyklos atstovų požiūris į kalbą buvo pagrįstas supratimu, kad žmonės civilizuotame pasaulyje turi bendrą minčių struktūrą[120] (Barsky 2017). Jų kalbos dalių nustatymas iš principo skyrėsi nuo antikinės gramatikos. *Tékhnē grammatikē* kalbos dalys buvo nustatomos pagal morfologinius požymius, Port Rojalio lingvistai rėmėsi semantiniais požymiais[121] (Rauh, 2010: 25). Pirmosios jų gramatikos *Grammaire générale et raisonnée* jau pats pavadinimas rodo, kad tai yra visuotinė gramatika, nesusijusi su konkrečios kalbos morfologija. Ji nustato tiesioginį ryšį tarp kalbos, suvokimo ir mąstymo, todėl yra universali. Tačiau šiam teiginiui pagrįsti pavyzdžiai buvo imami tik iš graikų, lotynų ir prancūzų kalbų (Rauh, 2010: 27–28). 1830 m. Vilhelmas fon Humboltas (Wilhelm von Humboldt) pasiūlė idėją, kad universali gramatika yra smulkiausių nedalomų gramatinių kategorijų ir jų santykių rinkinys, tarsi statybiniai blokai, iš kurių sudaromos sintaksinės struktūros, o

[117] “[...] categories are proposed for a particular language when they appear to be needed for that language, with little thought as to their applicability to the grammar of other languages” (Newmeyer 2008 iš Dąbrowska 2015).

[118] “[...] at this stage of linguistic inquiry almost every new language that comes under the microscope reveals unanticipated new features” (Evans, Levinson 2009).

[119] “[...] all languages are built upon a common grammar” (Nordquist 2018).

[120] “[...] the 17th century Port Royal grammarians, whose rationalist approach to language and language universals was based on the idea that humans in the “civilized world” share a common thought structure” (Barsky 2017).

[121] “[...] the famous Port Royal grammar, which provides a coherent and comprehensive analysis of language from the point of view of philosophy and logic and characterizes parts of speech as semantic categories” (Rauh 2010: 25).

vėliau joms uždedami apribojimai. Taip sukuriamos visų kalbų konkrečios gramatikos. Universalioje gramatikoje iškeliama prielaida, kad visos kalbos turi tą patį gramatinių kategorijų ir sintaksinių ryšių rinkinį, o žmonės, turėdami baigtinį šių priemonių skaičių, sudaro begalinį jų pavartojimo atvejų kiekį (Barsky 2017). Pastaruoju metu pasirodė daug publikacijų, bandančių paneigti šiuos teiginius (apie tai išsamiau žr. 5.4.2 poskyryje).

Universalios gramatikos idėją XX a. išpopuliarino Noamas Chomskis (Noam Chomsky). Tačiau jis pateikė kiek susiaurintą jos sąvoką, kuri apima tik baigtinį konkrečių kalbų skaičių[122] (Chomsky 1993: 13). Reikia pabrėžti, kad N. Chomskis padėjo pagrindus šiuo metu labai paplitusiam sakinių sintaksinės struktūros pavaizdavimo būdui – frazių gramatikai. Sakinį S (sentence) jis vaizdavo kaip sudarytą iš daiktavardinės frazės NP (noun phrase) ir veiksmažodinės frazės VP (verb phrase): S → NP INFL VP (Chomsky 1993: 25). Aiškindamas universalios gramatikos principus, kaip vieną jų jis nurodė veiksnio būtiną ar nebūtiną buvimą sakinyje. Universali gramatika atskiroms kalboms palieka galimybę pasirinkti, kuris atvejis atitinka jos poreikius. Anglų ir prancūzų kalbose veiksnys būtinas, tačiau kitose kalbose, pvz., semitų, nėra tokio reikalavimo, todėl universalioje gramatikoje šis parametras žymimas skliaustuose: S → (NP) INFL VP (Chomsky 1993: 27), t. y. parametro NP paėmimas į skliaustus rodo, kad jis gali būti sakinyje arba jo gali nebūti.

Sukurta universalios gramatikos teorija iš principo yra tik pasakojimas apie visus galimus kalbos garsus, leksinius konceptus, lingvistines reikšmes. Ji turi apimti visus galimus fonologinius ir semantinius požymius, visas taisykles ir apribojimus, kad būtų galima juos sujungti į žodžius, o žodžius – į begalinį skaičių frazių ir sakinių. Žinoma, tokia sudėtinga teorija niekada negalės būti išbaigta, bet šiuo požiūriu lingvistikos padėtis nėra prastesnė už chemijos, fizikos ar kitų mokslo sričių. Jose darbai taip pat nėra užbaigti[123] (McGilvray 2018).

Šiuo metu kalbų skaičius pasaulyje – nuo 5 000 iki 8 000 ir tik mažiau nei 10 proc. iš jų, t. y. apie 500, yra tinkamai dokumentuotos, t. y. turi išsamiai aprašytas gramatikas bei žodynus. Ir tai yra viskas, iš ko galima daryti apibendrinimus. Kai kurie

[122] "[...] UG [Universal Grammar] does in fact permit only a finite number of core [particular] grammars" (Chomsky 1993).

[123] Of course, such a complete theory may never be fully achieved, but in this respect, linguistics is no worse off than physics, chemistry, or any other science. They too are incomplete" (McGilvray 2018).

mokslininkai teigia, kad istorijos pradžioje kalbų galėjo būti apie pusę milijono, taigi, manytina, kad turime tik apie 2 proc. visos kalbų įvairovės (Evans, Levinson 2009: 432).

5.4.2. Morfologinių kategorijų ribos

Pastaruoju metu universali gramatika sulaukia nemažai kritikos. Pasigirsta net labai kategoriškų teiginių – kad universali gramatika iš viso neegzistuoja[124] (McCrum 2012). Kritikuojama plačiai paplitusi prielaida, kad visos kalbos yra panašios į anglų kalbą, tik skiriasi savo garsų sistema ir žodynu. Teigiama, kad jos skiriasi iš esmės visuose savo struktūros lygmenyse ir kad yra itin mažai universalijų, kurios būtų būdingos visoms kalboms (Evans, Levinson 2009: 429).

Pagrindinis ir neginčijamas faktas, susijęs su kalbomis, yra jų įvairovė, pvz., kalbos gali turėti mažiau nei 12 skirtingų garsų ir gali turėti jų 12 × 12, o gestų kalbose iš viso nenaudojami garsai. Kalbos gali neturėti morfologinės darybos, o jų semantika gali skaidyti pasaulį labai skirtingais pjūviais. Sintaksinė kalbų struktūra gali skirtis tiek, kiek parodyta 117 pav. (t. y. kiek gali skirtis žodžių tvarka įvairiose kalbose), ir tiek, kiek – 118 pav. (t. y. polisintetinėse kalbose vienas žodis atitinka visą anglų kalbos sakinį).

This woman caught that huge butterfly

That$_{\text{object}}$ **this**$_{\text{subject}}$ **huge**$_{\text{object}}$ **caught woman**$_{\text{subject}}$ **butterfly**$_{\text{object}}$

117 pav. Kalbų sintaksės skirtumai (parengta pagal Evans, Levinson 2009: 431)

I cooked the wrong meat for them again.

abanyawoihwarrgahmarneganjginjeng

118 pav. Polisintetinių ir analitinių kalbų skirtumai (parengta pagal Evans, Levinson 2009: 432)

Izoliacinės kalbos neturi kaitybinių asmens, skaičiaus, laiko afiksų, jos naudoja tik šaknis (73 interneto nuoroda[125]), o daugiskaitą ar būtąjį laiką gauna arba iš konteksto, arba iš kitų nepriklausomų žodžių. Polisintetinės kalbos į vieną ilgą žodį sudeda maždaug tokį pat turinį, koks anglų kalboje išreiškiamas sakiniu (118 pav.). Lao

[124] Interview *Daniel Everett: 'There is no such thing as universal grammar'* (McCrum 2012).

[125] Prieiga internete: https://lt.wikipedia.org/wiki/Izoliacin%C4%97_kalba [žiūrėta 2022-11-22].

kalba neturi būdvardžių ir ypatybę išreiškia kaip veiksmažodžio potipį. Kai kuriose kalbose nėra skirtumo tarp veiksmažodžio ir daiktavardžio, pvz., jos turi tokius žodžius kaip *bėgti*, *būti žmogumi*, *būti dideliam* ir pan. (Evans, Levinson 2009: 434). Kai kurios kalbos neturi priemonių spalvoms, skaičiams išreikšti, pvz., piraha kalba. Įdomu tai, kad šios tautos atstovai nėra sukūrę mitų, piešinių ir kolektyvinė atmintis tesiekia dvi paskutines kartas, tačiau iki šių dienų tauta yra išlikusi vienakalbė (Everett 2005: 621).

Esant tokiai kalbų įvairovei, logiška, kad lingvistai ima ieškoti kitų būdų, kaip parašyti įvairias gramatikas, anglų kalbos nelaikant pagrindu visiems tyrimams, t. y. nesuabsoliutinant teiginio, kad daiktavardis, veiksmažodis ir būdvardis yra tarpkalbinės kategorijos, nes visos kalbos juos turi[126] (Baker 2003: 298). Martinas Haspelmatas (Martin Haspelmath) kritikuoja kalbininkus, bandančius savo darbuose aprašyti kokią nors kalbą, pvz., čamorų, naudojant anglų kalbos kategorijas, kurios laikomos universaliomis (Chung 2012), ir teigia, kad tai tas pat, kas bandyti aprašyti anglų kalbą naudojant čamorų gramatikos kategorijas: *klasė I*, kurią sudaro žodžiai, turintys objektus, ir *klasė II*, kuriai priklauso visi likę žodžiai, t. y. neturintys objektų (Haspelmath 2012: 121). Jis pateikia vaizdų palyginimą šia tema ir sako, kad negalima klausti, ar kalba X skiria daiktavardį ir veiksmažodį, ar kalba X skiria veiksmažodį ir būdvardį, ar kalba X skiria daiktavardį ir būdvardį. Tarpkalbiniame lygmenyje negali būti klausiama, ar visos kalbos skiria daiktavardį ir veiksmažodį, ar visos kalbos skiria veiksmažodį ir būdvardį, ar visos kalbos skiria daiktavardį ir būdvardį. Tokius klausimus jis palygina su klausimu, kiek valstijų sudaro Prancūziją. Šito galima klausti apie JAV, bet ne apie Prancūziją (Haspelmath 2012: 109–114). Taigi, daiktavardis, būdvardis ir veiksmažodis yra apibendrinimai, tinkantys tik atskiroms kalboms.

Reikia rasti universalius požymius, kurie nebūtų specifiniai kurios nors atskiros kalbos bruožai. Simonas Floidas (Simon Floyd), nurodydamas skirtumus tarp daiktavardžio ir būdvardžio kečujų kalboje, teigia, kad būdvardis sakinyje eina prieš daiktavardį, bet ne atvirkščiai (Floyd 2011: 53 iš Haspelmath 2012: 117). Toks požymis gali būti bendras su anglų kalba, bet jis tinka ne visoms kalboms, pvz., to negalima pasakyti apie lietuvių kalbą. Universalesnis teiginys būtų, kad būdvardžiai gali eiti daiktavardžio pažyminiu, bet daiktavardžiai negali atributiškai pažymėti būdvardžių[127] (Haspelmath 2012: 117). Tačiau pasigilinus ir šiam teiginiui galima rasti

[126] "[...] all languages have at least a few nouns, verbs and adjectives" (Baker 2003).

[127] "[...] nouns cannot attributively modify adjectives" (Haspelmath 2012).

jį paneigiančių pavyzdžių: sakinyje *Kietas lauke ir skystas kambaryje gali būti tik vanduo žiemos metu* daiktavardžiai *lauke* ir *kambaryje* yra atitinkamai būdvardžių *kietas* ir *skystas* aplinkybiniai pažyminiai, t. y. pažymi juos atributiškai.

Metodologija nėra tiksli, jei lyginamos įvairių kalbų kategorijos, kurios atskirose kalbose nustatytos, remiantis skirtingais kriterijais. Ta pati kategorija gali būti apibrėžiama labai nevienodai, pvz.: senovės graikų Dionisijaus Trakiečio gramatikoje daiktavardis nusakomas kaip linksniais kaitoma kalbos dalis, reiškianti daiktą arba veiksmą; anglų kalboje daiktavardis – tai žodis, kuris gali eiti po apibrėžiklių *the*, *this*, *that*; kinų mandarinų kalboje (angl. *mandarin Chinese*) daiktavardis apibrėžiamas kaip žodis, kuris eina po klasifikatoriaus.

Kalbas reikia lyginti naudojant specialių konceptų rinkinį, pvz., vienas tokių sprendimų buvo lyginti kalbas, remiantis semantinėmis sąvokomis: *daiktų šaknys*, kurios reiškia fizinius objektus; *veiksmų šaknys*, kurios reiškia valingą veiksmą; ir *požymių šaknys*, kurios reiškia savybę. Bet kurioje kalboje galima lengvai nustatyti šaknis (o ne žodžius!) ir taip pat lengvai galima nustatyti daiktus, veiksmus ir savybes (Haspelmath 2012: 122–123).

Kalbos gali būti lyginamos remiantis šaknų ryšiu su pasakymais. Anglų kalbos daiktavardžio šaknies *water*, veiksmažodžio šaknies *run* ir būdvardžio šaknies *wet* ryšį su trimis pagrindiniais pasakymo tipais – paminėjimu, predikacija ir pažymėjimu – galima pailiustruoti lentele (119 pav.).

	Paminėjimas	**Predikacija**	**Pažymėjimas**
Daiktų šaknys	WATER	(that) **is** water	(colour) **of** water
Veiksmų šaknys	**the** runn-**ing**	(it) RUN (s)	runn-**ing** (water)
Požymių šaknys	the wet-**ness**	water **is** wet	WET (water)

119 pav. Šaknų ryšiai su pasakymais (parengta pagal Haspelmath 2012: 124)

Paprastai kalbose galima pastebėti tokias tendencijas: jei daikto šaknis paminima, ji neturi jokio funkciją nurodančio kodavimo, t. y. nominalizavimo; jei veiksmo šaknis pateikiama kaip predikatas, ji neturi specialaus funkcijos kodavimo, t. y. jungties; jei požymio šaknis naudojama kaip pažyminys, ji neturi specialaus funkciją nurodančio kodavimo, pvz., posesyvinio linksnio ir pan. Tačiau kečujų kalboje daiktų šaknys pažymėjimo atveju naudojamos taip pat, kaip ir požymių šaknys, o tagalų kalboje visų trijų rūšių šaknys naudojamos vienodai visuose trijuose pasakymo

tipuose. Taigi, absoliučiai universalių kriterijų, tinkančių palyginti visas kalbas, kol kas nepasiūlyta. Tačiau, jeigu tarp kalbų nebūtų nieko bendra, nebūtų įmanomas joks vertimas; reikia tik gebėti parinkti lygiagrečias formas visuose teksto sluoksniuose (Mockus 2018). 2022 m. gegužės mėn. duomenimis *Google* galėjo atlikti vertimus tarp 133 pasaulio kalbų[128] (Caswell 2022).

5.5. Lietuvių kalbos dalis *Gramatinėje struktūroje*

Net pačios moderniausios automatinio vertimo sistemos, veikiančios statistiniais ir neuroninių tinklų metodais, kol kas dar negali pakeisti žmogaus darbo, todėl dalis mokslininkų atlieka tyrimus ir taisyklėmis pagrįsto automatinio vertimo srityje. Šiuo metu, verčiant nedidelės apimties tekstus, taisyklėmis pagrįstas metodas duoda geresnius rezultatus nei tikimybiniai metodai. Tačiau problema yra žodžių apimtis: jų kiekiui didėjant, vertimo kokybė labai krinta. Todėl ieškoma būdų, kaip išspręsti šią problemą. Atliekant vertimus, paremtus skaitmeninėmis gramatikomis, tikrinami du pagrindiniai dalykai:

1) ar vertimo metu gautas rezultatas yra gramatiškai taisyklingas sakinys (žodžių junginys) toje kalboje, į kurią verčiama;
2) ar vertimo metu gautas rezultatas turi tą pačią reikšmę, kaip ir pateiktas versti sakinys (žodžių junginys).

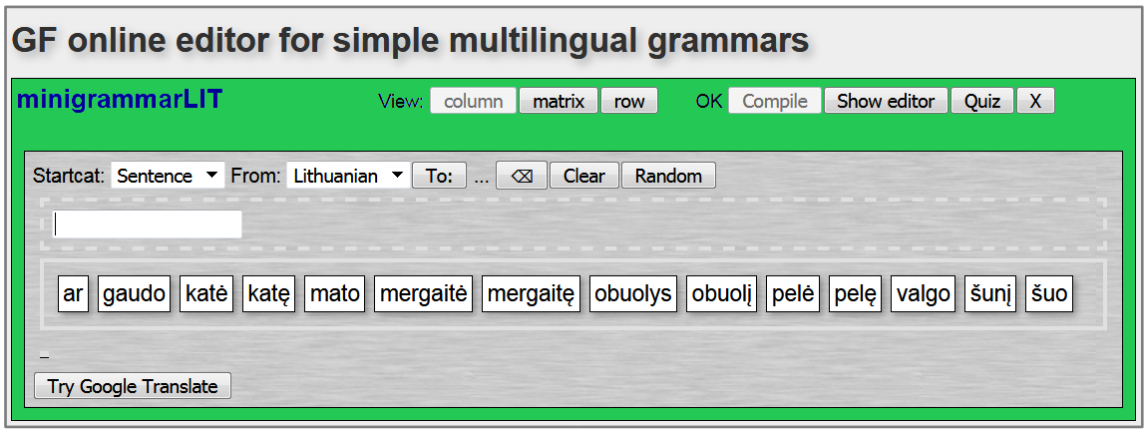
Pastaruoju metu išpopuliarėjęs neuroninių tinklų metodas, naudojamas automatinio vertimo sistemose, palyginus su statistiniu vertimu, turi ir privalumų, ir trūkumų. Privalumai: pagerėjo vidutinė vertimo kokybė ir sakiniai tapo sklandesni. Trūkumai: vertimo sistemos darbą yra daug sunkiau paaiškinti ir dar sudėtingiau numatyti vertimo rezultatus. Skaitmeninių gramatikų biblioteka *Gramateka* sukurta taip, kad leistinomis laikytų tik gramatiškai taisyklingas konstrukcijas (Ranta 2014: 3).

2017 m. vasaros kursų Rygoje metu buvo sukurtas pirmas bandomasis lietuvių kalbos skaitmeninės gramatikos pavyzdys *minigrammarLIT*, kuris prieinamas internete (74 interneto nuoroda[129]). Iš pradžių atliekant eksperimentą buvo pasirinkta nuostata atskleisti esminius skirtumus tarp anglų ir lietuvių kalbų: laisva / griežta žodžių tvarka

[128] "...languages to Translate, now supporting a total of 133 used around the globe".

[129] Prieiga internete: http://cloud.grammaticalframework.org/gfse/ [žiūrėta 2022-11-22].

ir artikelių buvimas / nebuvimas. Nutarta skirtingą žodžių tvarką lietuvių kalbos sakiniuose pasistengti atspindėti skirtingais artikeliais: lietuvių kalbos SVO sakiniams anglų kalbos papildinį naudoti su nežymimuoju artikeliu, pvz., *Mergaitė valgo obuolį – The girl eats an apple*, nes tai neutrali žodžių tvarka. Naudojant SOV žodžių tvarką, lietuvių kalboje pabrėžiamas papildinys, todėl į anglų kalbą nutarta tokį sakinį versti papildiniui naudojant žymimąjį artikelį: *Mergaitė obuolį valgo – The girl eats the apple* ir pan. Bandomasis pavyzdys apima visus galimus žodžių tvarkos variantus lietuvių kalbos sakiniuose. Žodyną tesudaro penki daiktavardžiai: *katė*, *mergaitė*, *obuolys*, *pelė* ir *šuo*; trys veiksmažodžiai: *gaudyti*, *matyti*, *valgyti*; ir klausiamasis žodelis *ar*. Kaip veikia sukurtas bandomasis pavyzdys, galima matyti pasirinkus skirtuko *minibar* / *show editor* variantą *minibar* tinklalapyje *Gramatinė struktūra* (74 interneto nuoroda[130]), o pasirinkus variantą *show editor,* pateikiamas programinės įrangos kodas. Atsidariusiame lange yra trys skirtukai: *Abstract*, *English*, *Lithuanian*, nes šiame bandomajame pavyzdyje numatyti vertimai tarp dviejų kalbų. Skirtuke *Abstract* užkoduota abstrakti sintaksė, bendra abiem kalboms, t. y. užkoduotos sąvokos. Visas skaitmeninės gramatikos aprašas sudaromas anglų kalba. Kalbų skirtukuose pateikiami sąvokų atitikmenys konkrečios kalbos žodžiais. Skirtuko *minibar* pradinis langas pateiktas 120 pav. Jame matyti visi žodžiai, kurie įtraukti į bandomąjį pavyzdį. Pažymėjus norimą žodį, jis perkeliamas į sakinio laukelį. Taip surenkamas visas sakinys. Kai žodžių rinkinį laukelyje, pvz., *katė mato pelę*, sistema atpažįsta kaip galimą sakinį, pateikia jo vertimą, tiksliau, abstrakčios sintaksės sąvokas, sakinio struktūrą (šiuo atveju SVO) ir sakinius abiem kalbomis (121 pav.).

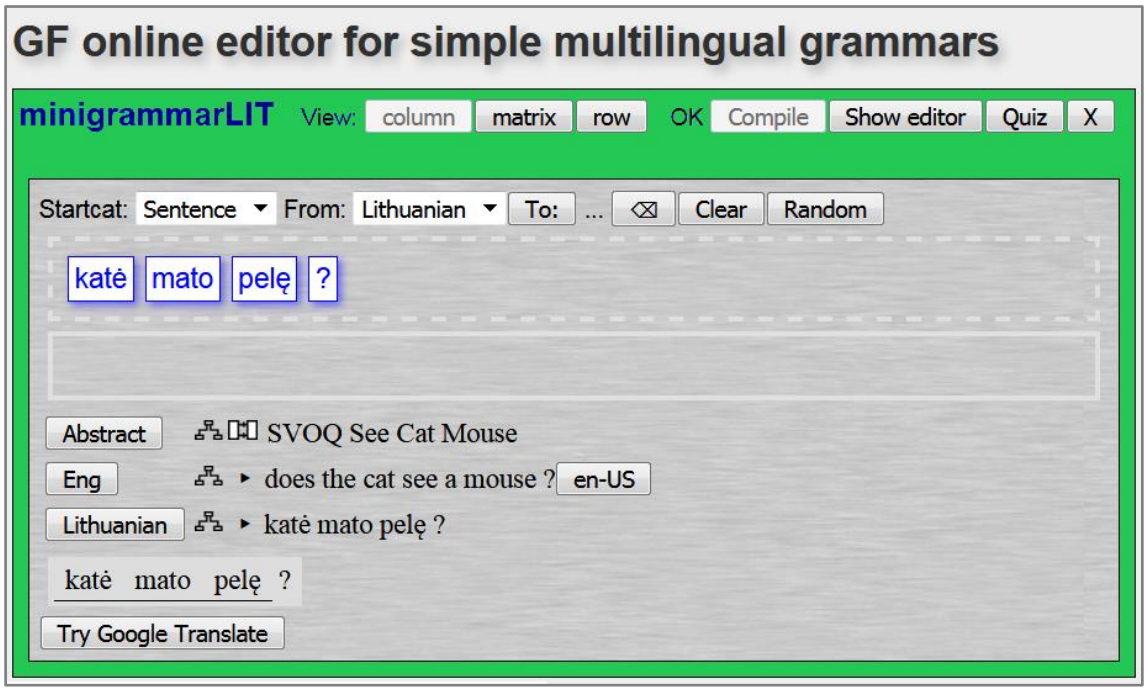


120 pav. Skirtuko *minibar* / *show editor* pradinis langas pasirinkus variantą *minibar*

[130] Prieiga internete: http://cloud.grammaticalframework.org/gfse/ [žiūrėta 2022-11-22].

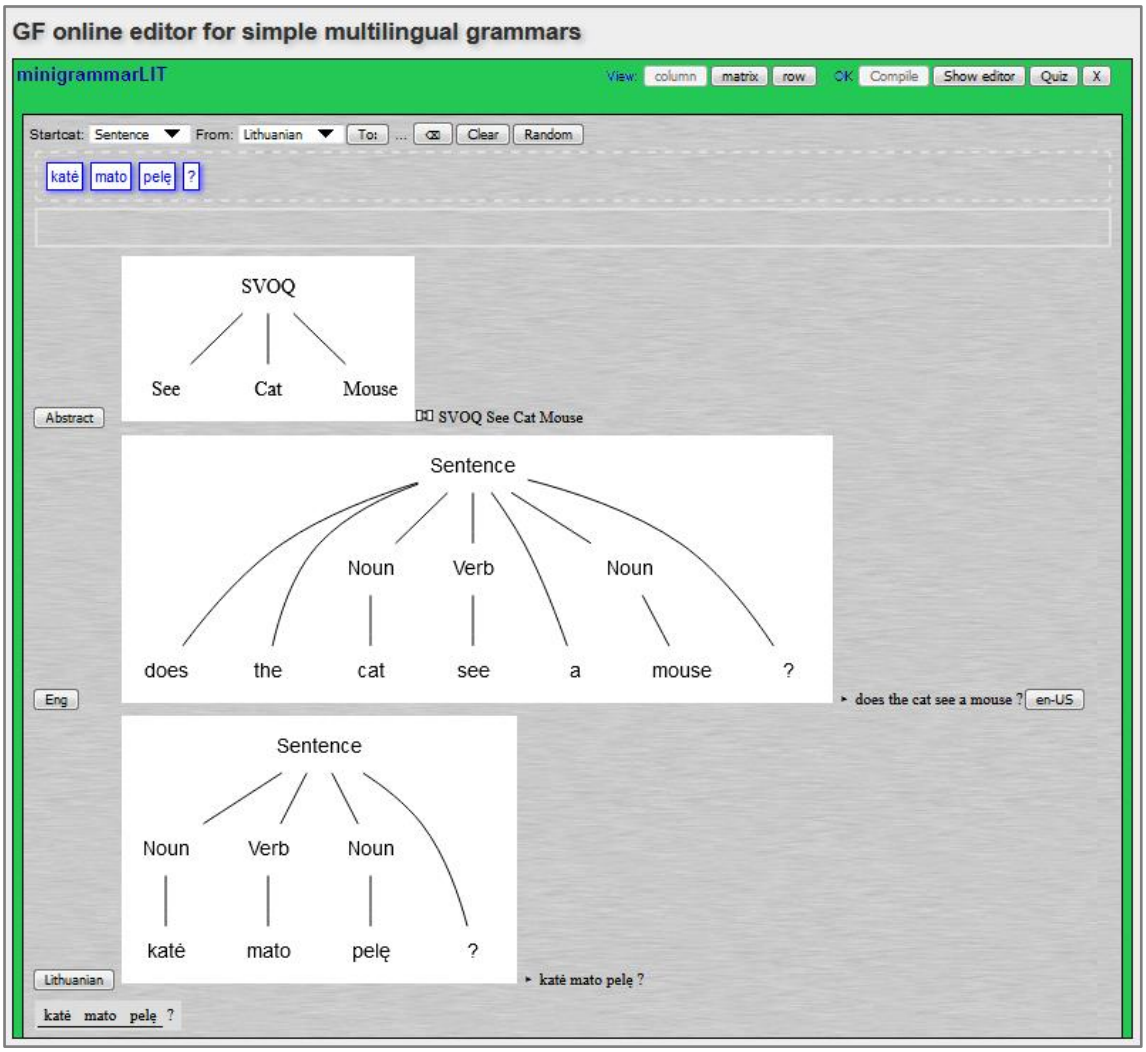
121 pav. Sakinys *Katė mato pelę*

Prie šių žodžių pridėjus klaustuką, anglų kalbos sakinys jau bus kitas, kaip ir sakinio struktūra abstrakčioje dalyje, kuri dabar jau yra SOVQ (122 pav.). Raidė Q, papildžiusi prieš tai buvusią struktūrą (SVO), rodo, kad tai yra klausiamasis sakinys.



122 pav. Sakinys *Katė mato pelę?*

Pažymėjus ikonas, esančias šalia mygtukų *Abstract*, *Eng*, *Lithuanian*, galima pamatyti sintaksinių struktūrų medžius – tiek konkrečios, tiek abstrakčios sintaksės. 122 pav. pateikto sakinio struktūra parodyta 123 pav.

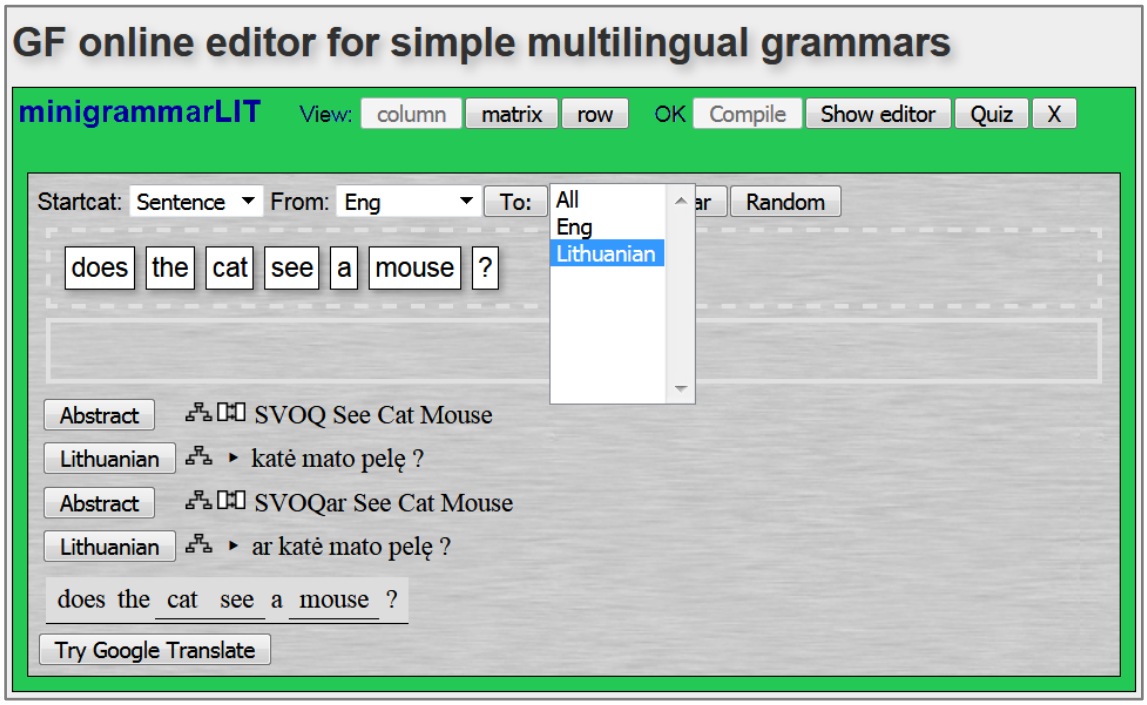


123 pav. Sakinys *Katė mato pelę?* su sintaksės medžiais

Tokia pati angliško sakinio *Does the cat see a mouse?* (123 pav.) sintaksinė struktūra gaunama ir tuo atveju, kai verčiamas lietuvių kalbos sakinys *Ar katė mato pelę?* (124 pav.). Šiame paveikslėlyje gerai matyti abstrakčios sintaksės sąvokų laukelyje (apatinis laukelis kairėje) pabraukti žodžiai, kurie atitinka aprašytas sąvokas. Klaustukas ir žodis *ar* nepriklauso sąvokoms, jie yra tarnybiniai simboliai, todėl lieka nepabraukti.

124 pav. Sakinys *Ar katė mato pelę?*

Atliekant vertimus iš anglų kalbos į lietuvių kalbą, sakiniui *Does the cat see a mouse* pateikiami abu galimi lietuviško sakinio variantai – tiek abstrakčios sintaksės lygmenyje (sakinio struktūros SVOQ ir SVOQar), tiek žodinėje sakinio išraiškoje (125 pav.).



125 pav. Sakinys *Does the cat see a mouse?*

Čia taip pat labai aiškiai matyti pabrauktos abstrakčios sintaksės sąvokos; joms nepriklauso tarnybiniai žodžiai *does*, *the*, *a* ir klaustukas, todėl jie lieka nepabraukti.

Kaip jau minėta, kuriant šį bandomąjį pavyzdį buvo stengtasi rasti skirtingus angliškų sakinių atitikmenis kiekvienam galimam lietuviško sakinio žodžių tvarkos variantui. VOS tipo sakiniams versti pasirinkta anglų kalbos konstrukcija *there is*. OVS tipo sakiniams naudoti pasyvo konstrukcijas, pvz., sakinio *Pelę mato katė* vertimą pateikti kaip *A mouse is seen by the cat*, taip atskiriant šią žodžių tvarką turintį sakinį nuo SVO sakinio *Katė mato pelę – The cat sees a mouse*. Tačiau visada išlaikomas sąvokų tapatumas: išverstame sakinyje yra visos originalo kalboje pavartotos sąvokos ir neatsiranda jokių naujų prasminių žodžių, kurių nebuvo originalo sakinyje.

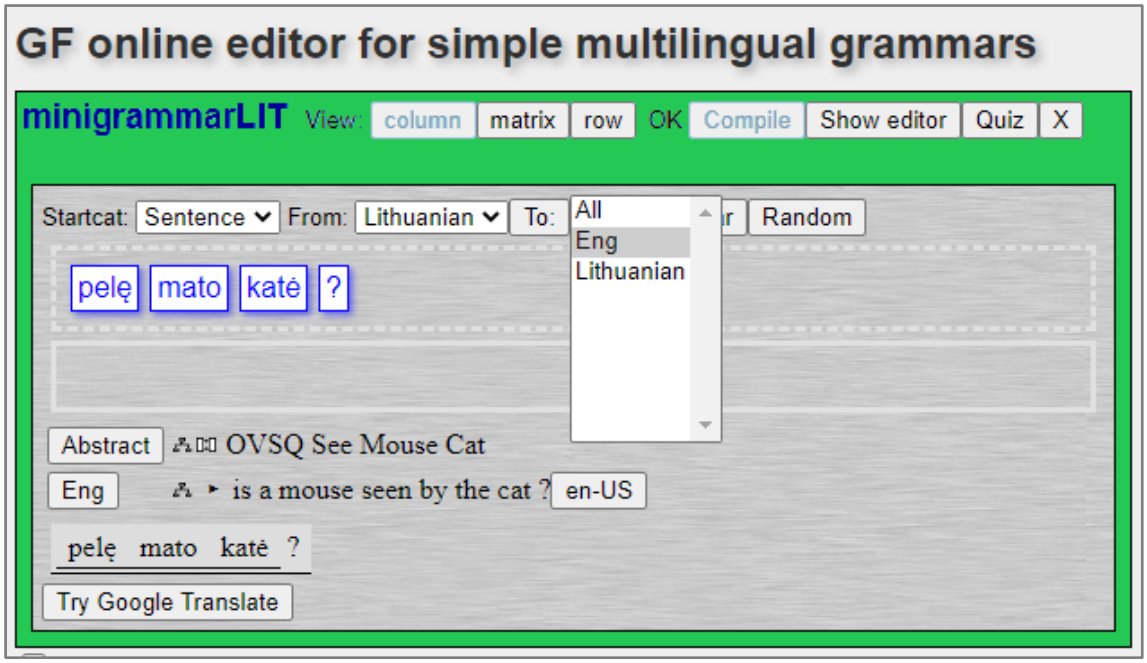
Tai, žinoma, nėra galutinis ir nekintamas vertimo tipas. Jei ateityje lietuvių kalbos skaitmeninė gramatika bus toliau vystoma, bus tariamasi su vertėjais ir atsižvelgiama į jų pasiūlymus bei pastabas. Tačiau net ir naudojant šias konstrukcijas (*there is* ar pasyvines konstrukcijas) skaitmeninė gramatika leidžia vertimo metu išlaikyti nepakitusias sąvokas, ko negalima pasakyti apie *Google* vertimus.

Taigi, pabaigoje belieka palyginti vertimus, atliekamus pasitelkus lietuvių kalbos skaitmeninę gramatiką, su *Google* vertimais, atliekamais naudojant statistinius metodus. Lietuvių kalbos sakinius, kuriuose yra anglų kalbos žodžių tvarka, *Google* sistema išverčia visai gerai. Netikslus vertimas prasideda tada, kai lietuviško sakinio žodžių tvarka nebeatitinka anglų kalbos žodžių išsidėstymo sakinyje. 126 pav. ir 127 pav. pavaizduotas sakinys *Pelę mato katė?* išverstas į anglų kalbą atitinkamai skaitmenine gramatika (74 interneto nuoroda[131]) ir *Google* vertimo sistema (75 interneto nuoroda[132]).

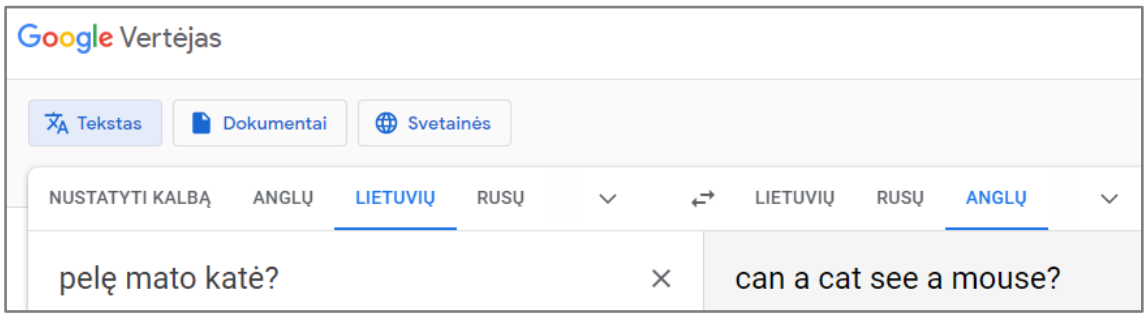
Google neišlaiko sąvokų tapatumo: katais žodžiai išnyksta, o kartais atsiranda naujos, lietuviškame sakinyje nesančios sąvokos. Žodžio *gali*, kuris panaudotas angliškame vertime (*can)*, originalo sakinyje nebuvo, t. y. žodžių vertimas neatitinka lietuvių kalbos sakinyje pavartotų sąvokų.

[131] Prieiga internete: http://cloud.grammaticalframework.org/gfse/ [žiūrėta 2022-11-22].

[132] Prieiga internete: https://translate.google.com/?sl=lt&tl=en&text=pel%C4%99%20mato%20kat%C4%97%3F&op=translate [žiūrėta 2022-11-22].



126 pav. Sakinio *Pelę mato katė?* vertimas, atliktas remiantis skaitmenine gramatika



127 pav. Sakinio *Pelę mato katė?* vertimas naudojant *Google* vertimo sistemą

Padėka. Lietuvių kalbos skaitmeninės gramatikos bandomasis pavyzdys buvo parengtas vasaros kursų *Grammatical Framework*, vykusių Rygoje 2017 m., metu. Dviem jų dalyviams, lietuvių kalbos skaitmeninės gramatikos bandomojo pavyzdžio autoriams, buvo skirta Latvijos valstybinė stipendija. Latvių kalbos skaitmeninė gramatika jau yra sukurta. Šiuo metu neskiriama finansinių išteklių lietuvių kalbos skaitmeninei gramatikai sukurti, tačiau siektina, kad ir Lietuvoje šiuo klausimu situacija keistųsi.

5.6. Skyriaus išvados

Šiuo metu terminu *skaitmeninė gramatika* įvardijami trijų pagrindinių sričių dalykai. Pirma, tai elektronine forma išleisti įvairių kalbų gramatikos vadovėliai. Antra, nedidelės apimties, greit sukuriamos apžvalginės gramatikos, pateikiančios tik esmines žinias apie kalbą. Tokios gramatikos dažniausiai reikalingos, kai kalba apdorojama kompiuteriu statistiniais metodais. Bene tiksliausiai šio termino esmę atspindi trečia jo alternatyva – formalus gramatikos taisyklių aprašas, naudojamas kuriant programinę įrangą dažniausiai automatinio vertimo tikslams. Vienas tokių pavyzdžių – *Gramatinė struktūra*, skirta skaitmeninėms gramatikoms rašyti. Visos skaitmeninės gramatikos saugomos *Gramatekoje* – skaitmeninių gramatikų bibliotekoje. Šiuo metu jos jau yra sukurtos 38 kalboms, tarp jų estų ir latvių, tačiau lietuvių kalbos šiame sąraše dar nėra.

Bet kokia programinė įranga, apdorojanti vienu metu daugelio tautų kalbas, turi būti iš karto kuriama kaip labai universali sistema. Idėjų sukurti universalią gramatiką galima įžvelgti dar viduramžiais, tačiau iki šiol nepavyko sukurti gramatikos, kuri tiktų visoms pasaulio kalboms. Kalbos iš esmės skiriasi visuose savo struktūros lygmenyse. Šiuo metu kritikuojama plačiai paplitusi prielaida, kad visos jos yra panašios į anglų kalbą. Polisintetinės kalbos į vieną ilgą žodį sudeda maždaug tokį pat turinį, koks anglų kalboje išreiškiamas visu sakiniu.

Lietuvių kalbai kol kas sukurta tik nedidelė apžvalginė gramatika ir skaitmeninės gramatikos bandomasis pavyzdys.

6. *LIETUVIŲ KALBOS GRAMATIKOS INFORMACINĖ SISTEMA* (LIGIS)

Gramatikos informacinė sistema – tai tarsi popieriuje spausdintų gramatikų inversija. Paprastai gramatikos vadovėliuose pateikiamos taisyklės, kurios tinka tam tikrai žodžių grupei, bet apsiribojama vien keliais pavyzdžiais ir neišvardijami visi žodžiai, vartojami pagal tą taisyklę, pvz., *Dabartinėje lietuvių kalbos gramatikoje* apibrėžiant daiktavardžių giminę pavyzdžių pateikiama po keturis vyriškosios ir moteriškosios giminės žodžius (Ambrazas 1997: 62). Visiems kitiems gramatikoje nepaminėtiems lietuvių kalbos žodžiams žmonės nesunkiai patys priskiria giminę. Kompiuteris pats nieko negali nuspręsti, jam turi būti labai tiksliai nurodyti visi kiekvieno žodžio gramatiniai duomenys, taigi, ir giminė. Todėl, kuriant gramatikos informacinę sistemą, į kalbą bandoma žiūrėti kitu aspektu: ne iš gramatinių kategorijų pozicijos, bet iš žodžio pozicijos, t. y. išeities taškas turi būti ne gramatikos taisyklė ir kaip jos iliustracija pateikti keli žodžiai, kuriems ji tinka, bet pats žodis turi būti pagrindas ir iš gramatikos išrenkami duomenys apie jį pagal visas su juo susijusias taisykles.

Esminis bruožas, skiriantis *Lietuvių kalbos gramatikos informacinę sistemą* (toliau – LIGIS) nuo kitų šiuo metu atliekamų lietuvių kalbos kompiuterizavimo darbų, – tai 100 proc. patikimumas, t. y. joje visiškai netoleruojamos klaidos. Būtent jau atliktų lietuvių kalbos gramatikos kompiuterizavimo darbų daromos klaidos, tiksliau, jau sukurtų automatinės morfologinės analizės sistemų nepakankamas tikslumas ir paskatino pradėti kurti gramatikos informacinę sistemą, pateikiančią išsamią ir labai tikslią informaciją. V. Zinkevičiaus sukurta programinė įranga lietuvių kalbos morfologinei analizei ir sintezei atlikti (Zinkevičius 2000) tobulai dirba tik sintezės dalyje. Analizuojant žodžius, daromos klaidos pradinės formos nustatymo metu, pvz., *padaryti* nurodoma, kad tai žodžio *padarytis* šauksmininko linksnis. Ši programinė įranga buvo naudojama kuriant taisyklėmis pagrįstą lietuvių kalbos automatinės sintaksinės analizės sistemą. Morfologinė analizė joje yra pirmasis etapas (Šveikauskienė 2010: 86), todėl tokios klaidos jau iš pat pradžių nulemdavo netikslų sistemos darbo rezultatą. Taigi, iškilo poreikis turėti 100 proc. patikimą morfologinę informaciją apie žodį.

6.1. Gramatinės informacijos rūšys ir jos pateikimas

LIGIS duomenys saugomi dviem lygmenimis – tai 1) plačiajai visuomenei skirta, labai populiariai ir visiems suprantamai pateikiama informacija ir 2) kompiuteriniam kalbos apdorojimui tinkantis formatas. Į sistemą įtraukiami tik dabartinės bendrinės lietuvių kalbos žodžiai.

6.1.1. Plačiajai visuomenei skirta informacija

Siekiant kuo patogiau vartotojams pateikti informaciją, buvo naudojamas prisitaikantis prie ekrano dydžio dizainas (angl. *Responsive Web Design* – RWD), kad sistema LIGIS (76 interneto nuoroda[133]) būtų galima naudotis ir mobiliuosiuose telefonuose. Tam tikslui informacija buvo išdėstyta į korteles (angl. *tiles*).

6.1.1.1. Duomenų bazė

Pagrindinis gramatikos kompiuterizavimo tikslas – elektronine forma turėti visą informaciją apie kiekvieną žodį. Morfologinę informaciją galima sukaupti dviem būdais: sudaryti formalios gramatikos taisyklių rinkinį, kuriuo naudodamasis kompiuteris kiekvieną kartą sugeneruotų žmogui reikalingus duomenis, arba surašyti į kompiuterio atmintį informaciją apie visų žodžių visas formas, kad kompiuteris prireikus galėtų ją iš ten perskaityti ir pateikti vartotojui. Pirmojo būdo buvo atsisakyta dėl kelių priežasčių. Kad kompiuteris galėtų savarankiškai apdoroti žodžius, jam turi būti labai tiksliai nurodyta, ką su kiekvienu žodžiu jis turi padaryti. Iš pradžių visus žodžius reikia taip suskirstyti į grupes, kad visiems tam tikros grupės žodžiams būtų galima taikyti tas pačias taisykles. Kiekvienai grupei turi būti nustatyta daug požymių, pagal kuriuos žodžiai bus jai priskiriami. Pavyzdžiui, atliekant automatinę lietuvių kalbos sintaksinę analizę, naudojamas laiko požymis: sakiniuose *Visą naktį ji skaitė tą knygą* ir *Visą knygą ji perskaitė tą naktį* tik dėl šio (laiko) požymio galima teisingai nustatyti sakinio dalis. Žodžiai *knygą* ir *naktį* morfologiškai nesiskiria niekuo: abu yra moteriškosios giminės vienaskaitos galininko linksnio. Ir tik laiko požymis leidžia

[133] Prieiga internete: http://ligis.lki.lt/ [žiūrėta 2022-11-22].

nustatyti, kad *naktį* yra aplinkybė, o *knygą* – papildinys (Šveikauskienė 2010: 95). Kažką panašaus reikėtų padaryti ir morfologijos srityje.

Viena pirmojo lietuvių kalbos morfologinio analizatoriaus daromų klaidų yra ta, kad kartais neteisingai nustatoma pradinė forma (lema). Kadangi dauguma darybos būdų buvo įtraukti kaip reguliarūs, t. y. būdingi visiems žodžiams, neatsižvelgiant į jų semantiką, todėl gauta daug nerealių homonimų, t. y. lietuvių kalboje neegzistuojančių žodžių, kurių tam tikros formos sutampa su esamais lietuvių kalbos žodžiais, pvz.: *gulėjas*, *neišskubėjas* ir pan. (Rimkutė 2006: 38), nes jų kilmininko linksnio forma sutampa su trečiojo asmens veiksmažodžiu (*gulėjo*, *neišskubėjo*). Veiksmažodinių daiktavardžių vediniai su priesaga *-ėj-* laikomi leistinais lietuvių kalboje žodžiais net ir tuo atveju, kai šaknis šios priesagos prisijungti negali.

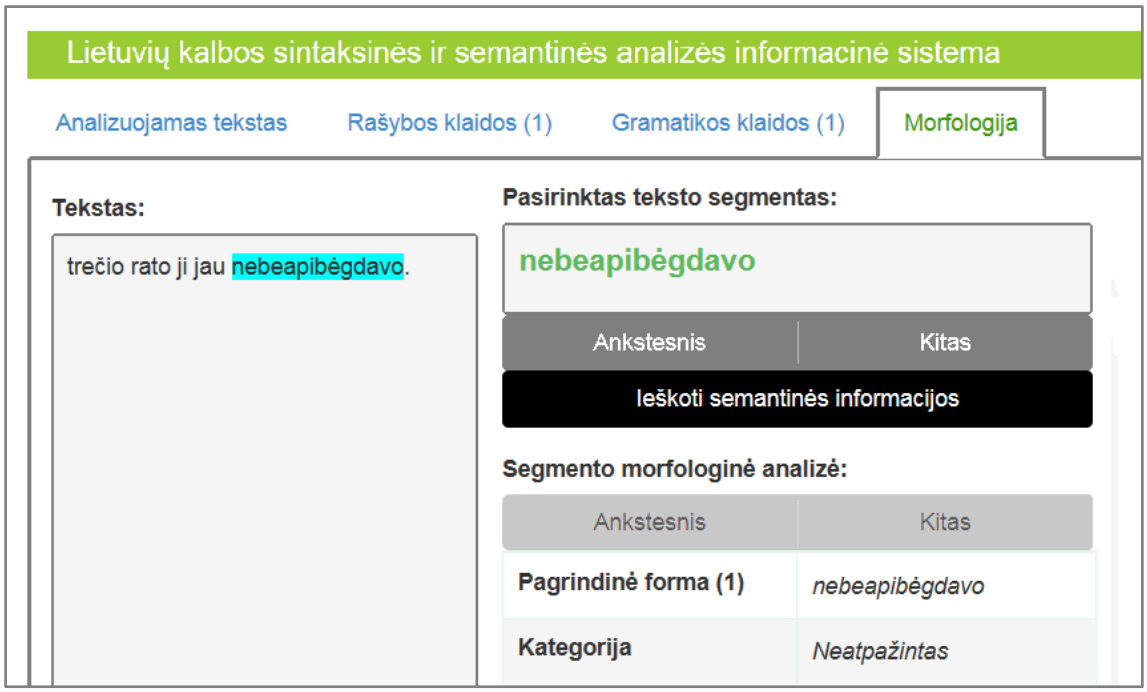
Taigi, reikėtų nustatyti požymius, kurie leistų kompiuteriui nuspręsti, kada žodis gali turėti priesagą *-ėj-* ir kada – ne. Tačiau tai padaryti labai sunku. Kokį bendrą požymį turi, pvz., žodžiai *gerovė*, *žinovas* ir *daržovė*, kad jie visi trys gali prisijungti priesagą *-ov-*, ir kas juos skiria nuo kitų žodžių, kurie šios priesagos negali turėti? Juk negalima pasakyti **kėdovė*[134]. Ir koks požymis tai rodo? Todėl, kuriant LIGIS, nuspręsta eiti antruoju keliu: kaupti kompiuterio atmintyje išsamią morfologinę informaciją apie visų žodžių visas formas.

Svarbiausia nuostata kuriant LIGIS buvo išvengti dviejų kraštutinumų, pasitaikančių jau atliktuose lietuvių kalbos kompiuterizavimo darbuose: perteklinių, lietuvių kalboje neegzistuojančių žodžių pateikimo ir per mažos žodžių apimties. Žodynuose nepateikti visi galimi dariniai, pvz., žodis *nebeatsinešti*. Kompiuteriniai kalbos apdorojimo įrankiai taip pat neapėmė daugelio žodžių. VDU Kompiuterinės lingvistikos centre sukurta *Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema* neatpažindavo retesnių žodžių, pvz., žodžio *nebeapibėgdavo* (128 pav.).

Autoriai, tyrinėję žodžių formas tekstyne, nurodo, kad jame pavartota labai nedaug kaitybinių formų: pagrindiniai vardažodžių linksniai yra vardininkas, kilmininkas ir galininkas (Rimkutė 2006: 39). Tekstyno pagrindu parengtoje morfemikos duomenų bazėje sukaupta apie 75 000 įrašų, tačiau su šaknimi *bėg-* yra tik 118 žodžių, pvz., dalyvio *bėgantis* tėra dvi vienaskaitos formos: kilmininkas ir galininkas (Šveikauskienė, Ribikauskas, Šveikauskas 2017: 150), taigi, net jo lema, t. y.

[134] Žvaigždute žymimi negalimi žodžiai.

vardininko linksnis, neįtraukta. Palyginimui galima pasakyti, kad LIGIS šaknies *bėg-* žodžių skaičius (su visomis jų kaitybinėmis formomis) yra apie 28 000.



128 pav. *Lietuvių kalbos sintaksinės ir semantinės analizės informacinės sistemos* pateikti duomenys apie žodį *nebeapibėgdavo* (43 interneto nuoroda, žiūrėta 2018-04-10)

Dėl perteklinių formų galima pateikti pavyzdį iš tinklalapio *morfologija.lt*. Jame žodžiui *susitikimas* nurodoma ne tik daiktavardžio forma, bet ir būdvardžio bei dalyvio (77 interneto nuoroda[135]), t. y. tai, kas nėra vartojama lietuvių kalboje. Pateikiama net įvardžiuotinė jo forma *susitikimoji*. 129 pav. parodyta žodžio *susitikimas* analizė. *Dabartinės lietuvių kalbos žodyne* žodžiui *susitikimas* nurodomas tik daiktavardžio variantas (Keinys ir kt. 1993: 778) ir net neįmanoma įsivaizduoti, kokio daiktavardžio pažyminiu galėtų būti toks būdvardis. Akademiniame *Lietuvių kalbos žodyne* (Naktinienė ir kt. 2017) prie žodžio *susitikimas* taip pat neparašyta, kad tai gali būti ir būdvardis. Net ir pagal lietuvių kalbos gramatikos taisykles būdvardžių vediniai su priesaga *-imas* galimi tik iš būdvardžių, ir tik tokių, kurių pamatiniai žodžiai retai bevartojami, pvz.: *artimas*, *gretimas*, *svetimas* (Ambrazas 1997: 201). Tritomėje gramatikoje taip pat pateikiami būdvardžių vediniai su priesaga *-imas* tik iš būdvardžių: *artimas*, *tolimas* ir kt. (Ulvydas 1965: 556).

[135] Prieiga internete: https://morfologija.lietuviuzodynas.lt/zodzio-formos/susitikimas [žiūrėta 2022-11-22].

129 pav. Žodžio *susitikimas* analizė tinklalapyje *morfologija.lt* (77 interneto nuoroda)

Stengiantis į LIGIS įtraukti visus lietuvių kalboje esančius žodžius, taigi, ir vedinius, kurie nepateko į žodynus, kompiuteriu buvo generuojamos visos priešdėlių ir šaknų bei priesagų kombinacijos. Vėliau gautos lemos buvo peržiūrėtos kalbininkų ir ranka atmesti lietuvių kalboje neegzistuojantys žodžiai. Kaitybinės formos generuojamos automatiškai, panaudojant V. Zinkevičiaus sukurtą lietuvių kalbos morfologinės sintezės programinę įrangą (Zinkevičius 2000), kurios darbe klaidų nebuvo pastebėta, todėl manoma, kad ši programinė įranga dirba 100 proc. patikimumu.

LKI buvo atlikti išsamūs lietuvių kalbos priešdėlių tyrimai (Šveikauskienė 2015a) ir nustatyta, kad yra apie 600 galimų priešdėlių kombinacijų, pvz.: *te-be-per-par-* (*tebeperparduodavo*), *ne-be-ati-* (*nebeatitolsta*) ir kt. Paaiškėjo, kad tiktai apie 220 iš jų gali būti vartojami su veiksmažodžiais. Kiti, pvz., *nuo-* (*nuomonė*), vartojami su daiktavardžiais ir pan.

6.1.1.2. Apibendrintas žodžio formatas

Kad būtų lengviau struktūriškai aprašyti morfologinius ir morfeminius žodžio duomenis, buvo sudarytas apibendrintas lietuvių kalbos žodžio formatas, apimantis visus galimus lietuvių kalbos žodžių struktūros variantus. Kiekvienas į duomenų bazę įtraukiamas žodis talpinamas į apibendrintą formatą.

Pirmojoje pakopoje lietuvių kalbos žodis skaidomas į priešdėlių, šaknų, priesagų ir galūnės sritis. Sudurtinių žodžių, turinčių tris ar keturias šaknis, lietuvių kalboje yra labai nedaug: *sienlaikraštis*, *dvdešimtpenkmetis* ir kt. Galūnė žodyje būna tik

viena arba žodis gali iš viso jos neturėti, pvz.: *aš*, *prie* ir kt. Dalelytiniai priešdėliai visada išsidėsto žodžio pradžioje ir niekada nebūna įsiterpę tarp prielinksninių priešdėlių (Šveikauskienė 2015a: 196). Dalelytinių priešdėlių žodyje negali būti daugiau kaip trys (*te-ne-be-at-si-neš-a*). Taip pat nebuvo pastebėta žodžių, turinčių daugiau nei tris prielinksninius priešdėlius.

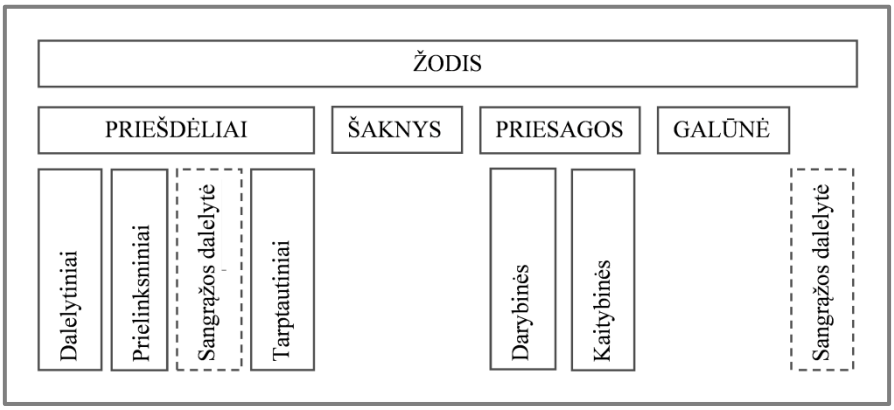
Lietuvių kalbotyroje įvairiai vertinami tarptautiniai priešdėliai. Šalia teiginių, kad „dabartinėje lietuvių kalboje yra vartojama tarptautinių veiksmažodžių su nelietuviškos kilmės priešdėliais [...]. Šie veiksmažodžiai traktuojami kaip nepriešdėliniai“ (Ulvydas 1971: 294), yra išsakytas ir kitas požiūris: „LKG II 294 visi tarptautiniai veiksmažodžiai rekomenduojami laikyti nepriešdėliniais. Morfeminei analizei toks požiūris vargu ar bepriimtinas“ (Kuosienė 1986: 65). Panašiai mano ir kiti mokslininkai: jei tas pats veiksmažodis yra vartojamas lietuvių kalboje ir be priešdėlio, pvz., *organizuoti* ir *reorganizuoti*, tai yra pagrindo *de-*, *re-* ir kt. laikyti atskiromis morfemomis, t. y. priešdėliais. Kita vertus, jei veiksmažodis gauna sangrąžos afiksą, jo negalima įterpti tarp šių priešdėlių ir šaknies (Pakerys 2014: 11).

Apie vardažodžius pasakyta: „Vardažodžiuose tarptautinius priešdėlius laikome priešdėliais tik tais atvejais, kai juos galima išskirti sugretinus su atitinkamais lietuvių kalboje vartojamais tarptautiniais ar lietuviškais nepriešdėliniais žodžiais“ (Kuosienė 1986: 65).

Akademinėje *Lietuvių kalbos gramatikoje* teigiama: „Dabartinėje lietuvių kalboje yra vartojami ne tik lietuviški, bet ir tarptautiniai priešdėliai“ (Ulvydas 1965: 435) ir pateikiama daug pavyzdžių, kuriuose jie prisišlieję prie lietuviškos šaknies žodžių, pvz.: *antifašistas*, *antimedžiaga*, *antidalelė*, *infraraudonieji (spinduliai)*, *kontekstas*, *rekristalizacija*, *subnuoma*, *subtropikai*, *ultragarsas* ir kt. (Ulvydas 1965: 436). Taigi, lietuvių kalboje skiriami trijų tipų priešdėliai: dalelytiniai, prielinksniniai ir tarptautiniai.

Dabartinės lietuvių kalbos gramatikoje rašoma, kad „sangrąžos formantas *si* visada eina tarp priešdėlio ir šaknies“ (Ambrazas 1997: 283). Tačiau tarptautiniai priešdėliai eina po dalelytės *si*, pvz., *nebesusikoncentruoja*. Jei tarptautinio priešdėlio nebelieka, dalelytė *si* rašoma prieš šaknį – *išsicentravo*.

Lietuvių kalboje sangrąžos dalelytė *si* gali būti prieš šaknį, t. y. priešdėlių srityje, arba žodžio gale. Kadangi ji žodyje gali būti tik viena, o apibendrintame žodžio formate jai skirtos dvi pozicijos, todėl abi jos pažymėtos punktyrine linija. Apibendrintas lietuvių kalbos žodžio formatas pateikiamas 130 pav.



130 pav. Apibendrintas lietuvių kalbos žodžio formatas

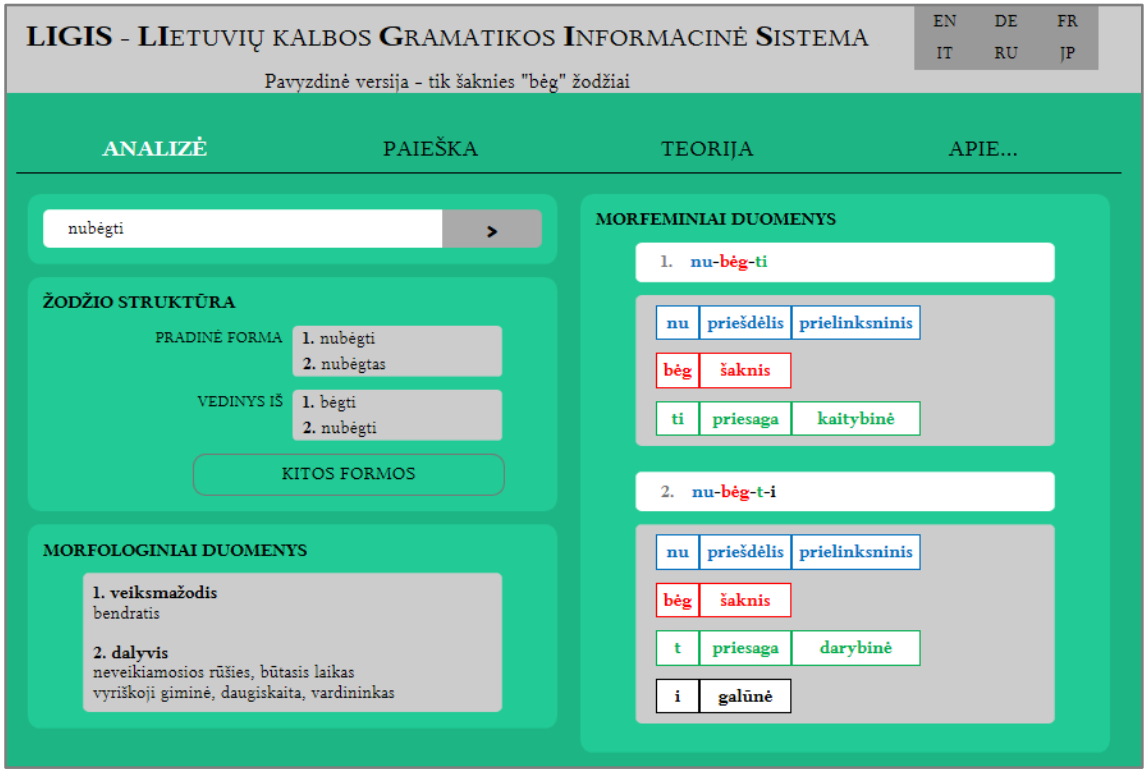
LIGIS apie kiekvieną morfemą nurodoma dar ir papildoma informacija, ne tik jos pavadinimas, pvz., jei šaknis turi intarpą ar balsių kaitą, tai atsispindės ir duomenų bazėje. Jei galūnė bus įvardžiuotinė ar sutrumpėjusi, tai taip pat bus pateikiama žodžio apraše.

6.1.1.3. Informacijos pateikimas internete

Pagrindinis bruožas, skiriantis LIGIS nuo kitų lietuvių kalbos kompiuterizavimo darbų, yra tas, kad viename tinklalapyje pateikiama trijų tipų informacija apie žodį: morfologinė, morfeminė ir darybinė. Kiekvienam žodžiui nurodoma pradinė forma, o vediniams ir dūriniams – dar ir pamatiniai žodžiai. Morfologinė informacija (kalbos dalis, giminė, skaičius, linksnis, laikas, asmuo ir pan.) pateikiama ne santrumpomis, o ištisais žodžiais ar žodžių junginiais. Apie morfeminę dalį reikia pasakyti, kad tai yra pirmas viešai prieinamas šaltinis, kuriame nurodomas morfemos tipas (šaknis, priesaga, galūnė ir t. t.) ir jos charakteristikos (priesaga – darybinė, kaitybinė; galūnė – įvardžiuotinė, sutrumpėjusi ir kt.). Kiekviena morfema žymima vis kita spalva: priešdėlis – mėlynai, šaknis – raudonai, priesaga – žaliai, galūnė – juodai, sangrąžos dalelytė *si* – rudai, daugiašaknių žodžių jungiamasis balsis – violetine spalva.

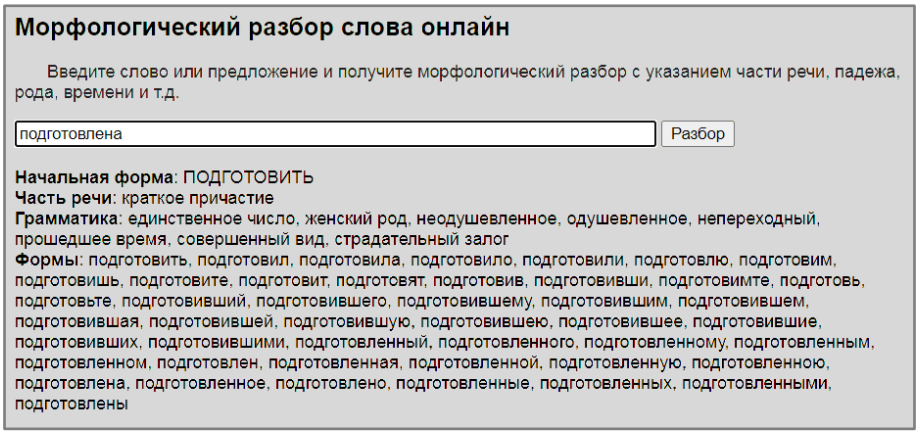
Jei žodis gramatiškai daugiareikšmis, lemos sunumeruojamos ir vėliau kiekvienai reikšmei pateikiami duomenys atitinkamai pagal reikšmės numerį. Gramatiškai daugiareikšmio žodžio *nubėgti* analizės pavyzdys pateiktas 131 pav. (78 interneto nuoroda[136]).

[136] Prieiga internete: http://ligis.lki.lt/?wordInput=nub%C4%97gti [žiūrėta 2022-11-22].



131 pav. Žodžio *nubėgti* analizės pavyzdys (78 interneto nuoroda)

Sistemoje numatyta galimybė peržiūrėti visas vartotoją dominančio žodžio kaitybines formas. Palyginimui galima pateikti rusų kalbos morfologinės analizės pavyzdį. Čia kiekvieną kartą įvedus žodį šalia morfologinės informacijos surašomos visos jo kaitybinės formos. 132 pav. parodytas žodžio *подготовлена* pavyzdys (79 interneto nuoroda[137]).



132 pav. Žodžio *подготовлена* morfologinės analizės pavyzdys (79 interneto nuoroda)

[137] Prieiga internete: https://goldlit.ru/component/slog?words=%D0%BF%D0%BE%D0%B4%D0%B3%D0%BE%D1%82%D0%BE%D0%B2%D0%BB%D0%B5%D0%BD%D0%B0 [žiūrėta 2022-11-22].

LIGIS visos formos pateikiamos atskirame lange, kuris atsidaro vartotojo pageidavimu pele paspaudus mygtuką KITOS FORMOS. Žodžio *bėgtakis* kaitybinės formos parodytos 133 pav. (80 interneto nuoroda[138]).

KITOS FORMOS

BĖGTAKIS

DAIKTAVARDIS, BENDRINIS, VYRIŠKOJI GIMINĖ

	Vienaskaita	Daugiskaita
Vardininkas	bėgtakis	bėgtakiai
Kilmininkas	bėgtakio	bėgtakių
Naudininkas	bėgtakiui	bėgtakiams
Galininkas	bėgtakį	bėgtakius
Įnagininkas	bėgtakiu	bėgtakiais
Vietininkas	bėgtakyje	bėgtakiuose
Šauksmininkas	bėgtaki	bėgtakiai

133 pav. Žodžio *bėgtakis* kaitybinės formos (80 interneto nuoroda)

Šiuo metu tik du žodžiai – *bėgis* ir *bėgti* – iliustruoti visų kaitybinių formų vartojimo pavyzdžiais. Ateityje numatoma jų parinkti kiekvienam sistemoje esančiam žodžiui. Žodžio *bėgį* vartojimo pavyzdžiai parodyti 134 pav. (81 interneto nuoroda[139]).

Visa informacija šiuo metu svetainėje pateikiama septyniomis kalbomis: lietuvių, anglų, vokiečių, prancūzų, italų, rusų ir japonų. Ateityje planuojama įtraukti dar tris kalbas: latvių, lenkų ir ispanų. Siekiant kuo didesnio tikslumo ir patikimumo, visa informacija, taigi, ir gramatikos terminai, buvo verčiami dviejų gimtakalbių: lietuvio, mokančio užsienio kalbą, ir tos kalbos lingvisto, mokančio kalbėti lietuviškai: į vokiečių kalbą buvo verčiama kartu su Krista Šnaider (Christa Schneider), į italų kalbą – su Adrianu Čeriu (Adriano Cerri), į japonų kalbą vertė Simona Vasilevskytė su Kajako Takagi ir t. t.

[138] Prieiga internete: http://ligis.lki.lt/?wordInput=b%C4%97gtakis [žiūrėta 2022-11-22].

[139] Prieiga internete: http://ligis.lki.lt/?wordInput=b%C4%97g%C4%AF [žiūrėta 2022-11-22].

134 pav. Žodžio *bėgį* vartojimo pavyzdžiai (81 interneto nuoroda)

Užsieniečiams, besimokantiems lietuvių kalbos, labai patogu, kad, pereinant į kitą kalbą, sistema įsimena vartotojo įvestą žodį, ir žmogus iškart gali matyti visą gramatinę informaciją ta kalba, kurią jis geriau moka, jei kai kurie lietuviški terminai jam dar nežinomi.

6.1.2. Gramatinių požymių kodavimas

Informacijos pateikimo forma, kuri patogi žmonėms, yra visai netinkama kompiuteriui, apdorojančiam kalbą, todėl gramatiniai požymiai koduojami.

Tekstai pradėti koduoti atsiradus tekstynams (Marcinkevičienė 2000: 12). Italas Robertas Buza (Roberto Busa) 1946 m. pasiūlė JAV firmai IBM kompiuteriu išanalizuoti Tomo Akviniečio tekstus. Darbai buvo pradėti po trejų metų, specialiai šiam tikslui sukūrus naujas technologijas, labiau pritaikytas žodžiams, o ne skaičiams

apdoroti (82 interneto nuoroda[140]). Kadangi tekstynai pradėti rinkti iš anglų kalbos tekstų, todėl ir žymos buvo kuriamos tokios, kad kuo geriau atspindėtų anglų kalbos savybes. Vėliau pradėti kaupti bei anotuoti ir kitų kalbų tekstynai. Įvairioms kalboms tiek pačios žymos, tiek jų kiekis labai skiriasi, nes kartais afiksais išreiškiama ta informacija, kuri anglų kalboje nusakoma žodžių tvarka ar tarnybiniais žodžiais, pvz., graikų kalbos žymų skaičius yra apie 1 200 (DeRose 1990). O polisintetinėse kalbose suteikti žodžiams žymas iš esmės nėra galimybės.

Automatiškai priskirti žodžiams žymas yra lengviau, kai jų skaičius mažesnis, todėl neseniai pasiūlyta universali sistema, apimanti tik 12 kategorijų, kurios yra bendros 22 kalboms (Petrov, Das, McDonald 2012: 2089). Žymų kiekis pasirenkamas atsižvelgiant į tekstyno panaudojimo tikslus. Daugiausia jų reikia atliekant kalbos analizę (dažniausiai statistinę) ir ieškant žodžių pavartojimo pavyzdžių. Anotuoti tekstynai leidžia atlikti gramatinių modelių paiešką, nenurodant konkretaus žodžio, pvz., surasti anglų kalbos sakinių pavyzdžius, kuriuose prieš daugiskaitos daiktavardį nebūtų artikelio; arba surasti žodį *help*, pavartotą kaip daiktavardį, po kurio eitų būtojo laiko veiksmažodis (83 interneto nuoroda[141]); arba pateikti lietuvių kalbos sakinių, kuriuose papildinio nederinamuoju pažyminiu eitų įvardis ir pan.

6.1.2.1. Pirmasis morfologinių požymių kodavimas Lietuvoje

Praeito amžiaus pabaigoje buvo sukurta morfologinės analizės ir sintezės programinė įranga, automatiškai nustatanti lietuvių kalbos žodžių morfologinius požymius. Jei vartotojas yra žmogus, rezultatai pateikiami naudojant sutrumpinimus: dažniausiai praleidžiamos balsės lietuviškuose morfologinių kategorijų pavadinimuose. 135 pav. matyti, kad žodžiui *valstybės* nurodoma kalbos dalis (daiktavardis), po kurios eina pradinė forma kampiniuose skliaustuose ir į atskiras eilutes išdėstyti visi galimi morfologiniai variantai (Zinkevičius 2000: 249). Tačiau kompiuterio vidiniame formate gramatinės kategorijos koduojamos skaičiais.

[140] Prieiga internete: https://web.archive.org/web/20120327122219/http://www.ibm.com/ibm100/it/en/stories/linguistica_computazionale.html [žiūrėta 2022-11-22].

[141] Prieiga internete: https://www.sketchengine.eu/blog/pos-tags/ [žiūrėta 2022-11-22].

```
valstybės
  dktv <valstybė>
    dktv mot.gim vnsk K
    dktv mot.gim dgsk V
    dktv mot.gim dgsk Š
```

135 pav. Žodžio *valstybės* morfologinė analizė (Zinkevičius 2000: 249)

Pirmasis morfologiškai anotuotas lietuvių kalbos tekstynas sukurtas daugiau nei prieš 20 metų ir jo pagrindu sudarytas *Dabartinės rašomosios lietuvių kalbos dažninis žodynas* (Grumadienė, Žilinskienė 1997). Jame kiekvienam žodžiui nurodyta kalbos dalis, koduojant ją trijų raidžių seka, pvz.: *dkt*, *bdv*, *įvr*, *vks*, *prv*, *dll* ir kt.

6.1.2.2. Žymų standartas baltų kalboms SGR

2015 m. buvo sukurtas bendras lietuvių ir latvių kalboms žymų rinkinys, besiremiantis LGR (Leipzig Glossing Rules). Pati idėja gimė ir pamatai jai buvo padėti vasaros kursų Salose metu, todėl ir pavadinimas, analogiškas Leipcigo žymų sistemai LGR (84 interneto nuoroda[142]), yra susijęs su šia Lietuvos vietove. Baltų kalbų koduotė vadinasi SGR – *Salos Glossing Rules* (Nau, Arkadiev 2015: 197).

Pagrindinė priežastis, paskatinusi sukurti SGR, buvo faktas, kad šiuo metu anotuojant žodžius vis dažniau žymos nurodomos kita kalba nei pats tekstas. Tokie darbai naudingi kalbininkams, besidomintiems gramatikomis, ir ypač tų kalbų, kurios jiems nėra žinomos. Bet kol kas dar nesukurti tvirti žymų standartai, tad mokslininkams painu analizuoti daugybę žymėjimo modelių, naudojamų dabartinėse publikacijose (Nau, Arkadiev 2015: 195).

Nors autoriai teigia, kad remiasi LGR, bet kai kur žymėjimai nesutampa, pvz., REFL keičiama į RFL. Toks sprendimas grindžiamas argumentu, kad netikslinga koduoti, pvz., neveikiamosios rūšies būtojo laiko dalyvį naudojant 13 simbolių (PST.PASS.PTCP), kai lietuvių kalboje šiai gramatinei kategorijai išreikšti naudojama tik viena raidė (Nau, Arkadiev 2015: 196). Taigi, kai kurie kodai trumpinami: vietoje PASS.PTCP naudojamas dviejų raidžių kodas PP ir pan. Neatitikimų galima pastebėti ne tik su LGR, bet ir su *Dabartine lietuvių kalbos gramatika* (Ambrazas 1997), pvz., SGR bendroji daiktavardžių giminė, atsižvelgiant į kontekstą, nurodoma kaip vyriškoji

[142] Prieiga internete: https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf [žiūrėta 2022-11-22].

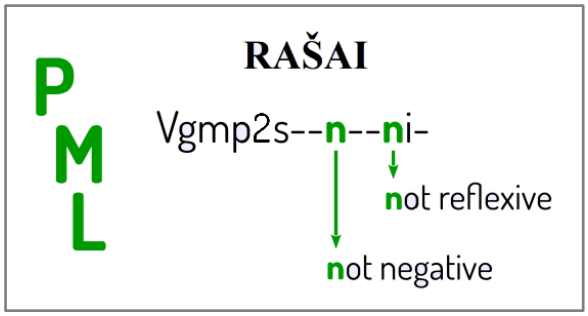
arba moteriškoji. Bendrosios giminės daiktavardžio *kvaiša (fool)* pavyzdys pateiktas 136 pav.

Lithuanian
a. *tok-s* *kvaiš-a*
such-NOM.SG.M fool-NOM.SG / fool(M)-NOM.SG
'such a fool' (referring to a man)
b. *toki-a* *kvaiš-a*
such-NOM.SG.F fool-NOM.SG / fool(F)-NOM.SG
'such a fool' (referring to a woman)

136 pav. Bendrosios giminės daiktavardžio kodavimas SGR (parengta pagal Nau, Arkadiev 2015: 204)

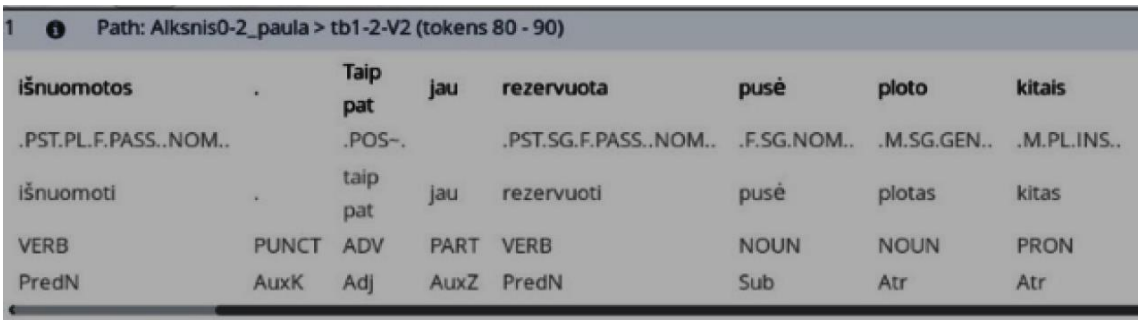
6.1.2.3. VDU morfologinio anotavimo formatai

VDU Kompiuterinės lingvistikos centre 2016 m. parengtas sintaksiškai anotuotas tekstynas ALKSNIS, kuriame sakinio struktūrai pavaizduoti naudojami du formatai: PML (Prague Markup Language), kurio pagrindą sudaro MULTEXT-East formatas, ir PAULA XML. PML privalumas yra tai, kad jis labai kompaktiškas, t. y. užima mažai vietos kompiuterio atmintyje. Tačiau, atliekant paiešką, jis nėra optimalus, nes reikia įvertinti ne tik paties simbolio reikšmę, bet dar ir poziciją, kurioje jis stovi. Tas pats simbolis skirtingose pozicijose gali turėti nevienodas reikšmes. Šio formato esmė būtų tokia: žodis pavaizduojamas simbolių seka, kurioje kiekvienas simbolis reiškia tam tikrą morfologinę kategoriją, pvz., žodžio *rašai* kodas yra *Vgmp2s--n--ni-* (*verb, general, main form, present tense, 2nd person, singular, no gender, no voice, not negative, no definiteness, no case, not reflexive, indicative mood, no degree*). Šiame žodžio *rašai* kode simbolis *n* devintoje pozicijoje turi reikšmę *not negative*, dvyliktoje pozicijoje – *not reflexive* (137 pav.).



137 pav. Žodžio *rašai* kodavimas PML

Todėl, siekiant pagerinti paieškos galimybes, imta naudoti dar vieną formatą – PAULA, kurio pagrindą sudaro LGR (Leipzig Glossing Rules). Čia panaikintas daugiareikšmiškumas, kiekvieną morfologinę kategoriją užkoduojant simbolių grupe, apribota iš abiejų pusių taškais. Tokio kodavimo pavyzdys (Bielinskienė ir kt. 2016: 110) parodytas 138 pav. Atliekant paiešką, pvz., požymį M (masculine gender) galima atskirti nuo NOM (nominative case), pateikus užklausą *.M.* (Bielinskienė ir kt. 2016: 111).



138 pav. Tekstyno fragmento [*patalpos jau*] *išnuomotos. Taip pat jau rezervuota pusė ploto kitais* [*metais iškilsiančiame statinyje.*] analizė (parengta pagal Bielinskienė ir kt. 2016: 110)

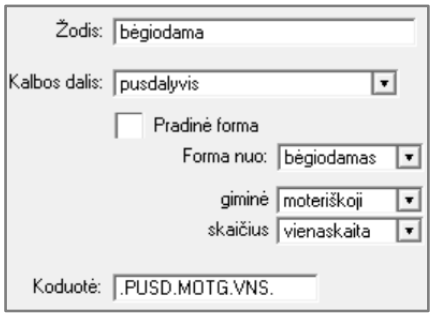
VDU kalbininkai PAULA žymas papildė lietuvių kalbai reikalingais, bet PAULA sistemoje nesančiais, terminais pridėdami bangelės ženklą prieš morfologinę kategoriją. Taip terminui *veikiamoji rūšis* (angl. *active voice*) buvo sukurtas kodas ~ACT. Bendroje latvių ir lietuvių kalboms koduotėje SGR jis žymimas kartu su dalyvio terminu: PA naudojama veikiamosios rūšies dalyviams, PP – neveikiamosios. Taigi, tas pats terminas įvairiose kodavimo sistemose žymimas skirtingai.

6.1.2.4. LIGIS žodžių morfologinės žymos

Pagrindinis LIGIS sukūrimo tikslas – paruošti lietuvių kalbos gramatikos dokumentaciją. Viena iš jos pritaikymo sričių – parengti lietuvių kalbos gramatikos kompiuterinį variantą, kuriame būtų patogu atlikti paiešką. Morfologijos dalyje turi būti sukaupta visa morfologinė informacija apie visas kiekvieno lietuvių kalbos žodžio formas. LIGIS plačiajai visuomenei skirtoje dalyje morfologiniai duomenys apie žodį pateikiami nenurodant linksniuotės ar asmenuotės – ši informacija būtų perteklinė, nes pateikiamos visos žodžio kaitybinės formos. Tačiau vidiniame kompiuterio formate, koduojant žodį, ji yra būtina. Atliekant mokslinius lietuvių kalbos tyrimus, svarbu ją turėti, pvz., nustatant, kiek tam tikros linksniuotės ar asmenuotės žodžių yra nagrinėjamame tekste ir pan.

Jau yra sukurti du LGR žymomis besiremiantys lietuvių kalbos kodavimo formatai: PAULA ir SGR. Jie abu skirti tekstams anotuoti. Kuriant PAULA buvo siekiama įtraukti lietuvių kalbą į tarptautinę mokslinių tyrimų infrastruktūrą CLARIN – *Common Language Resources and Technology Infrastructure Consortium* (Bielinskienė ir kt. 2016: 107). SGR tikslas – padaryti lietuvių kalbos anotuotus tekstus prieinamus jos nemokantiems kitų šalių tyrėjams (Nau, Arkadiev 2015: 197). Šie tikslai logiškai pateisina lietuvių kalbos tekstų anotavimą anglų kalbos kaip *lingua franca* gramatikos terminų žymomis. Tačiau net ir labai artimoms kalboms – latvių ir lietuvių – nebuvo lengva rasti bendrą, anglų kalbos gramatika paremtą kodavimą. Ten, kur šių abiejų baltų kalbų skirtumai itin žymūs, pasiūlyti skirtingi sprendimai (Nau, Arkadiev 2015: 195).

LIGIS svetainėje visa informacija pateikiama septyniomis kalbomis, tačiau duomenys nurodomi naudojant lietuviškų terminų vertimus į šias kalbas ir nebandoma lietuvių kalbos gramatikos įsprausti į anglų kalbos gramatikos rėmus. LIGIS žodžiams koduoti buvo pasirinktas PAULA kodavimo principas, tik naudojant lietuvių kalbos žodžiais nusakomų morfologinių požymių sutrumpinimus. Žodžio *bėgiodama* žymų formato pavyzdys parodytas 139 pav.



139 pav. Žodžio *bėgiodama* žymų formato pavyzdys sistemoje LIGIS

Tiesiogiai naudoti VDU sukurto lietuviškų žymų rinkinio JABLONSKIS negalima, nes jis neapima visų terminų, reikalingų LIGIS esančiai informacijai užkoduoti: ne visų kalbos dalių tipams yra suteikti kodai, pvz., nėra žymos jungtukams – sujungiamasis ar prijungiamasis; prielinksniams nenurodyta, su kokiu linksniu jie vartojami; kaitomiems žodžiams nenurodytos linksniuotės, asmenuotės ir kt. Kuriant patį žymų rinkinį JABLONSKIS taip pat nebuvo perimti jau naudoti 1997 m. pasirodžiusiame *Dažniniame dabartinės rašomosios lietuvių kalbos žodyne* (Grumadienė, Žilinskienė 1997) lietuviškų morfologinių kategorijų sutrumpinimai.

Žymų rinkinio formatu pateikta informacija apie žodį leidžia atlikti įvairialypę morfologinių duomenų paiešką. Tyrinėjantis lietuvių kalbą kitos šalies mokslininkas vis tiek privalėtų išsiaiškinti, ką reiškia kodas, kuris yra papildomai įtrauktas į LGR. Pateikiant kompiuteriui duomenis, kokios informacijos turi būti ieškoma, nėra didelio skirtumo, kurios kalbos santrumpą jam nurodyti – SG (singular) ar VNS (vienaskaita). Užklausą pateikiantis vartotojas vis tiek turi žinoti, ką tas kodas reiškia.

Sprendimas kurti savą, lietuvių kalbos savybes visiškai atitinkančią kodavimo sistemą nėra kažkuo ypatingas ar išskirtinis. Taip elgiasi ir kitų šalių mokslininkai, pvz., Latvijoje buvo sukurtas savas semantinis analizatorius (angl. *Abstract Meaning Representation Parser*) (Barzdiņš, Gosko 2016: 1143). Anotuodami tekstynus atskiri tyrinėtojai taip pat gali susikurti savus labai specializuotus žymų rinkinius, pritaikytus jų tyrimų reikmėms[143] (83 interneto nuoroda[144]).

6.2. Diskutuotini atvejai lietuvių kalbos gramatikoje: dalyvis

Pradėjus kurti LIGIS buvo domėtasi, kaip kalbos dalys aprašomos įvairiuose leidiniuose. Pastebėta daug nenuoseklumų aptariant dalyvį. Todėl šiame poskyryje bus apžvelgta, kaip buvo traktuojamas dalyvis nuo pirmųjų gramatikų iki šių dienų. Tai aptarti labai svarbu, nes LIGIS informacija apie dalyvį kaip kalbos dalį pateikiama kitaip nei akademinėje gramatikoje.

Reikia pasakyti, kad laikui bėgant keičiasi kalbininkų požiūris ne tik į dalyvį. Kartais ir tuo pačiu metu atskiri kalbininkai kai kuriuos kalbos reiškinius interpretuoja nevienodai. Ypač tai akivaizdu skirtingos paskirties leidiniuose, pvz., praeito amžiaus 8-ojo dešimtmečio pradžioje plačiajai visuomenei skirtose knygose bendraties morfema *-ti* buvo laikoma galūne. Juozas Žiugžda savo gramatikoje rašė: „Bendratis turi galūnę *-ti*“ (Žiugžda 1971: 139). Mokslinėse gramatikose tas pats morfologinis reiškinys vertinamas kitaip. Tų pačių metų laidos tritomėje lietuvių kalbos gramatikoje rašoma: „Bendratis turi priesagą *-ti* arba *-t*“ (Ulvydas 1971: 398). Vienatomėje lietuvių kalbos gramatikoje ši bendraties morfema taip pat vadinama priesaga (Ambrazas 1997: 383). Būtų galima rasti ir daugiau panašių pavyzdžių. Apžvelgiant XX a.

[143] “Individual researchers might even develop their own very specialized tagsets to accommodate their research needs” (83 interneto nuoroda).

[144] Prieiga internete: https://www.sketchengine.eu/blog/pos-tags/ [žiūrėta 2022-11-22].

gramatikas rašoma, kad J. Jablonskis pateikia tik du neveikiamosios rūšies būsimojo laiko dalyvius: *būsimas* ir *mylėsimas*. Jis teigia, kad šis dalyvis beveik visai išnykęs iš šnekamosios kalbos. Dabar liko tik *būsimas* (Paulauskienė 2015: 74). Tačiau šiandien vartojami ir kitų šaknų būsimojo laiko dalyviai. 2018 m. bėgimo maratono apraše pavartotas dalyvis *bėgsimas*. „Penktadienis: savaitgalį laukia didžiausias bėgsimas atstumas, bent jau kol kas.“ (85 interneto nuoroda[145]). Kitas pavyzdys: dalyvis *išleisimas* „Nors apie tikslų *išleisiamų* dainų kiekį R. Čivilytė kalba dar nedrąsiai, ji dalinasi, kad, baigiantis rudeniui, tikisi galėsianti koncertuose atlikti ne vieną singlą iš būsimo albumo.“ (86 interneto nuoroda[146]) Lietuvių kalbos instituto išleistoje kolektyvinėje monografijoje yra toks sakinys: „Be to, sovietiniu laikotarpiu *matysima* artimos erdvės dinamika ir Myerso (2008: 449) mintis, kad ‘svarbesnis už geografinį yra „funkcinis“ atstumas: kaip dažnai susikerta žmonių keliai’, leidžia autorėms presuponuoti, kad ...“ (Vaičiulytė-Semėnienė ir kt. 2022: 9). Ypač juntamas internete pastaruoju metu veikiamosios rūšies būsimojo laiko dalyvių vartojimo pagausėjimas, 10 priede pateikiama keletas pavyzdžių.

LIGIS dalyvis interpretuojamas kitaip nei dabartiniuose kalbotyros leidiniuose – jis laikomas atskira kalbos dalimi, o ne veiksmažodžio forma. Toliau bus pateikti argumentai, pagrindžiantys šį teiginį.

Atkreiptinas dėmesys į tai, kad žodžiai įvairiose kalbose į klases skirstomi labai nevienodai. Beveik visos kalbos turi daiktavardį ir veiksmažodį, ir ne tik indoeuropiečių kalbos, pvz., tamilų kalba, kuri priklauso dravidų kalbų grupei (87 interneto nuoroda[147]). Apie 500 metų pr. m. e. datuojamame veikale, kuriame aprašoma tamilų kalbos gramatika, žodžiai skirstomi į keturias grupes: 1) daiktavardžius, 2) veiksmažodžius, 3) žodžius, kurie rodo ryšį tarp veiksmažodžio ir daiktavardžio, ir 4) žodžius, kurie pažymi daiktavardį ar veiksmažodį (88 interneto nuoroda[148]).

---

[145] Prieiga internete: http://www.dovydas.sankauskas.lt/wiki/maratono_treniruot%C4%97s_pradedantiesiems_10_savaitė [žiūrėta 2018-03-03].

[146] Prieiga internete: https://www.lrytas.lt/zmones/muzika/2020/06/29/news/ivertinkite-pokycius-kardinaliai-stiliu-pakeitusi-rosita-civilyte-pristate-daina-man-gana--15441488/ [žiūrėta 2022-11-22].

[147] Prieiga internete: https://en.wikipedia.org/wiki/Tamil_language [žiūrėta 2022-11-22].

[148] Prieiga internete: https://en.wikipedia.org/wiki/Part_of_speech [žiūrėta 2022-11-22].

Tačiau, klasifikuojant kitus žodžius, esama didelių skirtumų. Valterio Jungo (Walter Jung) vokiečių kalbos gramatikoje skaitvardis nelaikomas atskira kalbos dalimi ir sakoma, kad žodžiai, kurie reiškia skaičius, gali būti būdvardžiai, daiktavardžiai, įvardžiai, prieveiksmiai[149] (Jung 1967: 359). Kitoje vokiečių kalbos gramatikoje skaitvardis priskiriamas būdvardžiui – vadinamas *Zahladjektiv* ir aprašomas būdvardžių skyriuje (Helbig, Buscha 1989: 320). Internete pateikiamas kompromisinis variantas, kad skaitvardis vokiečių kalbotyroje kartais laikomas savarankiška kalbos dalimi[150] (89 interneto nuoroda[151]), taigi, tik kartais, bet ne visada.

Dar vienas mums neįprastos žodžių charakteristikos pavyzdys galėtų būti japonų kalbos būdvardžiai. Jie kaitomi laikais (90 interneto nuoroda[152]), pvz.: *mėlynas – aoi*, *buvo mėlynas – aokatta*; *karštas – atsui*, *buvo karštas – atsukatta* ir pan.

Taigi, kiekvienos kalbos žodžiai į kalbos dalis skirstomi atsižvelgiant į specifines tos kalbos ypatybes. Akademinėje *Lietuvių kalbos gramatikoje*, išleistoje 1965 m., įrašytas toks teiginys: „Nereikia manyti, kad kalbos dalys pagal kokybę ir kiekybę gali būti vienodos atskirose kalbose“ (Ulvydas 1965: 31). Lietuvių autoriai, rašydami gramatikas, remiasi užsienio autorių darbais. Matyt, todėl ir dalyvis lietuvių kalbotyros darbuose laikomas veiksmažodžio forma. Tačiau yra argumentų, leidžiančių manyti, kad dalyvis galėtų būti traktuojamas kaip atskira kalbos dalis. Anglų kalboje dalyvis iš tikrųjų – tik forma, o lietuvių kalboje – visa paradigma. Mokykloms skirtuose gramatikos duomenų aprašuose ir plačiajai visuomenei pateikiamoje interneto informacijoje jau galima įžvelgti prielaidų dalyviui suteikti savarankiškos kalbos dalies statusą.

[149] „Das Numerale (das Zahlwort) ist keine Wortart im eigentlichen Sinne. Adjektive und Substantive, Pronomina und Adverbien können „Zahlwörter“ sein“ (Jung 1967: 359).

[150] „Zahlwort wird in der Sprachwissenschat manchmal als eigene Wortart angesetzt“ (89 interneto nuoroda).

[151] Prieiga internete: https://de.wikipedia.org/wiki/Zahlwort [žiūrėta 2022-11-22].

[152] Prieiga internete: https://en.wikipedia.org/wiki/Japanese_equivalents_of_adjectives [žiūrėta 2022-11-22].

6.2.1. Požiūrio į dalyvio sampratą raida

Šiuolaikiniuose indoeuropiečių prokalbės gramatikos tyrimuose teigiama, kad žodžiai skirstomi į aštuonias kalbos dalis: daiktavardžius, būdvardžius (įskaitant dalyvius), įvardžius, veiksmažodžius, prieveiksmius, prielinksnius, jungtukus ir jaustukus[153]. Dalyvių priskyrimas būdvardžiams aiškinamas tuo, kad jie, kaip ir būdvardžiai, parodo kokybę (Quiles, López-Menchero 2011: 153).

6.2.1.1. Dalyvio interpretavimas pirmosiose gramatikose

Seniausia išlikusi sanskrito gramatika, o kartu ir seniausia išlikusi gramatika pasaulyje, parašyta Paninio V a. pr. m. e. (91 interneto nuoroda[154]). Joje dalyvis dar neminimas. Dalyvio nėra ir Jaskos veikale apie sanskrito kalbą. Pats veikalas iki mūsų dienų neišliko, todėl galima spėti, kad buvo sukurtas dar anksčiau, nes Paninis jį cituoja (92 interneto nuoroda[155]).

Dionisijus Trakietis dar II a. pr. m. e. graikų kalba parašė seniausią žinomą antikinę gramatiką *Gramatikos mokslas* (*Tékhnē grammatikē*). Pagal jos modelį parašytos vėlesnės indoeuropiečių kalbų gramatikos (Kairienė 2003: 788). *Gramatikos moksle* skiriamos aštuonios kalbos dalys: vardažodis, veiksmažodis, dalyvis, artikelis, įvardis, prielinksnis, prieveiksmis, jungtukas (88 interneto nuoroda[156]). Apibūdindamas dalyvį jis rašė: „Dalyvis turi daiktavardžio ir veiksmažodio bruožų. Jis turi visus daiktavardžio kaitybos atvejus ir veiksmažodžio, išskyrus asmenį ir nuosaką“[157] (Rauh 2010: 17). Čia reikėtų atkreipti dėmesį, kad *Gramatikos moksle* kalbos dalys buvo nustatomos pagal morfologinius kaitybos požymius, o graikų kalbos daiktavardžių ir būdvardžių kaitybiniai požymiai tie patys (linksnis, skaičius, giminė), todėl nebuvo galimybių juos išskirti (Rauh 2010: 19).

---

153 “Words are divided into eight parts of speech: nouns, adjectives (including participles), pronouns, verbs, adverbs, prepositions, conjunctions, and interjections [...] A participle is a word that attributes quality like an adjective” (Quiles, López-Menchero 2011).

154 Prieiga internete: https://de.wikipedia.org/wiki/Panini_(Grammatiker) [žiūrėta 2022-11-22].

155 Prieiga internete: https://en.wikipedia.org/wiki/Y%C4%81ska [žiūrėta 2022-11-22].

156 Prieiga internete: https://en.wikipedia.org/wiki/Part_of_speech [žiūrėta 2022-11-22].

157 “Participle is a word partaking of the nature both of nouns and verbs. It has all the accidents which belong to nouns as well as those which belong to verbs, except mood and person” (Rauh 2010: 17).

Lotyniškos gramatikos buvo rašomos remiantis graikų kalbos gramatikomis. Varonas savo gramatikoje *De lingua latina* visus lotynų kalbos žodžius skirsto į kaitomus ir nekaitomus. Kaitomi žodžiai išskaidyti į keturias klases pagal du kriterijus – laikus ir linksnius: kaitomi ir laikais, ir linksniais (dalyviai); nekaitomi nei laikais, nei linksniais (prieveiksmiai); kaitomi tik laikais (veiksmažodžiai); kaitomi tik linksniais (vardažodžiai). Iš viso jis nustatė šešias kalbos dalis: 1) vardažodį, 2) veiksmažodį, 3) dalyvį, 4) prieveiksmį, 5) prielinksnį ir 6) jungtuką (93 interneto nuoroda[158]).

Nuo tradicijos laikyti dalyvį atskira kalbos dalimi nenukrypstama ir Kristupo Sapūno parašytoje (Teofilio Šulco išleistoje) lietuvių kalbos gramatikoje. Joje pateikiamos aštuonios kalbos dalys: keturios kaitomos (vardažodis, įvardis, dalyvis, veiksmažodis) ir keturios nekaitomos (prieveiksmis, prielinksnis, jungtukas ir jaustukas) (Eigminas, Stundžia 1997: 27). Danieliaus Kleino gramatikoje dalyvis aprašomas ne veiksmažodžio skyriuje: jis pateikiamas atskirame lygiaverčiame su kitomis kalbos dalimis skyriuje po veiksmažodžio prieš prieveiksmį (Balčikonis, Larinas, Kruopas 1957: 504), todėl yra pagrindo manyti, kad D. Kleinas dalyvį laikė atskira kalbos dalimi.

6.2.1.2. Informacijos apie dalyvį pateikimas dabartinėse gramatikose

Dabartinėse gramatikose (tiek senųjų kalbų, tiek šiuolaikinių) dalyvis aprašomas kaip veiksmažodžio forma. „Dalyviai, iš prigimties vardažodinės formos, ilgainiui prisišliejo prie veiksmažodžio formų sistemos“ (Ambrazas 1979: 16).

Dalyvis šiuolaikinėse senųjų kalbų gramatikose. Šiuolaikinėse senųjų kalbų (lotynų, graikų) gramatikose dalyvis nebelaikomas atskira kalbos dalimi. Jonas Dumčius *Trumpoje istorinėje graikų kalbos gramatikoje* dalyvį aptaria veiksmažodžio skyriuje kartu su kitomis veiksmažodžio morfologinėmis kategorijomis – laikais, asmenimis, skaičiais ir t. t. (Dumčius 2011: 100). Lotynų kalbos gramatikoje *Elementa Latina* dalyvis taip pat pateikiamas kaip neasmenuojamoji veiksmažodžio forma (Dumčius, Kuzavinis, Mironas 2010: 76).

[158] Prieiga internete: https://mokslai.lt/referatai/lietuviu-kalba/lingvistika.html [žiūrėta 2022-11-22].

Dalyvis šiuolaikinių indoeuropiečių kalbų gramatikose. Internete pateikiamoje informacijoje apie kalbos dalis pabrėžiama, kad anglų kalbos gramatika parašyta sekant europietiška tradicija, išskyrus tai, kad dalyvis laikomas veikiau veiksmažodžio forma, o ne atskira kalbos dalimi (88 interneto nuoroda[159]). Taip pateikta ir naujausiose anglų kalbos gramatikose. *Šiuolaikinėje anglų kalbos gramatikoje* (angl. *Grammar of Contemporary English*) dalyvis nurodomas veiksmažodžio formų lentelėje (Quirk ir kt. 1992: 62). *Oksfordo anglų kalbos gramatikoje* (angl. *The Oxford English Grammar*) dalyvis įvardijamas kaip nestandartinė veiksmažodžio forma (Greenbaum, 1996: 132). *Oksfordo anglų kalbos gramatikos vadovo* knygoje (angl. *Oxford Guide to English Grammar*) aprašomos aštuonios pagrindinės žodžių klasės: veiksmažodis, daiktavardis, būdvardis, prieveiksmis, prielinksnis, apibrėžiklis, įvardis ir jungtukas (Eastwood 2002: 1). Kaip matyti, dalyvio tarp jų nėra.

Vokiečių kalbos gramatikose dalyvis taip pat laikomas veiksmažodžio forma. Vienose jis vadinamas vardažodine veiksmažodžio forma, kuri apibūdinama kaip neasmenuojama[160] (Jung 1967: 202). Kitose jis nusakomas kaip beasmenė veiksmažodžio forma[161] (Helbig, Buscha 1989: 105). Naujausioje vokiečių kalbos gramatikoje *DEUGRA – eine Deutsche Grammatik*, kurioje jau pateikiami vokiški kalbos dalių bei morfologinių kategorijų pavadinimai vietoje iki šiol vartotų lotyniškų, pvz., dalyvis vadinamas terminu *Mittelwort* – pažodžiui verčiant *tarpinis žodis*, t. y. tarpinis tarp veiksmažodžio ir būdvardžio, tačiau jis vis tiek apibūdinamas kaip nežymimoji veiksmažodžio forma – *unbestimmte Verbform* (Paukert, Holböck 2017: 7).

Taip pat gramatikose vertinamas ir lietuvių kalbos dalyvis. Frydricho Kuršaičio gramatikoje, kuri buvo išleista 1876 m., dalyvio kirčiavimas aprašomas veiksmažodžių kirčiavimo skyriuje, taigi, galima teigti, kad dalyvį jis laikė veiksmažodžio forma, nors knygoje pasakyta, kad dalyvių kaitybos pagrindą sudaro būdvardžių linksniavimas (Kuršaitis 2013: 357), vadinasi, tam tikrą nenuoseklumą čia jau galima pastebėti. F. Kuršaičio gramatikoje matyti aiški vokiečių kalbos gramatikos įtaka. Dalyviai skiriami į būdvardiškuosius, nurodant vokiečių kalbos atitikmenį *Partizip I*, ir prieveiksmiškuosius – vokiečių kalbos atitikmuo *Partizip II* (Kuršaitis 2013: 358–359).

---

[159] Prieiga internete: https://en.wikipedia.org/wiki/Part_of_speech [žiūrėta 2022-11-22].

[160] „Nominalformen [...] heißen die Formen des Verbs, die nicht durch eine Person bestimmt sind“ (Jung 1967: 202).

[161] „infinite Verbform“ (Helbig, Buscha 1989: 105).

Veiksmažodžio forma dalyvis laikomas ir XX a. kalbotyroje. J. Žiugždos gramatikoje dalyvis apibrėžiamas kaip „veiksmažodžio forma, turinti veiksmažodžio ir būdvardžio ypatybių“ (Žiugžda 1971: 174). *Dabartinėje lietuvių kalbos gramatikoje* rašoma: „Dalyvis yra iš veiksmažodžio kamieno daroma linksniuojamoji forma, turinti veiksmažodžio ir būdvardžio ypatybių“ (Ambrazas 1997: 354).

Bene ilgiausiai senąją tradiciją – dalyvį laikyti savarankiška kalbos dalimi – išlaikė Rusijos kalbininkai. 2012 m. vadovėlyje mokykloms dalyvis aprašomas savarankiškų kalbos dalių skyriuje ir apibrėžiamas kaip savarankiška kalbos dalis, reiškianti daikto požymį, susijusį su veiksmu ir turintį būdvardžio bei veiksmažodžio ypatybių[162] (Бабайцева, Чеснокова 2012: 167). Nors yra ir rusų kalbos gramatikų, kuriose dalyvis apibrėžiamas kitaip: „Dalyvis – tai atributinė veiksmažodžio forma, jungianti dviejų kalbos dalių – veiksmažodžio ir būdvardžio – reikšmes“ (Шведова 1980: 664).

6.2.2. Dalyvio vieta lietuvių kalbotyroje

Lietuvių kalbos gramatikoje, kuri parašyta anglų kalba, dalyvis aptariamas neasmenuojamųjų veiksmažodžio formų[163] skyriuje (Ambrazas 2006: 284).

Akademinėje *Lietuvių kalbos gramatikoje* išvardijama vienuolika kalbos dalių, tačiau dalyvis tarp jų nepaminimas ir nurodoma, kad taip žodžiai buvo klasifikuojami ir J. Jablonskio gramatikoje, skyrėsi tik dalelytės ir ištiktuko vertinimas (Ulvydas 1965: 31).

2011 m. sudarytame *Dažniniame lietuvių kalbos morfemikos žodyne* apie dalyvį *bėgančio* pateikiama tokia informacija (140 pav.): pradinė forma *bėgti*; veiksmažodis, dalyvis, vyriškoji giminė, esamasis laikas, vienaskaitos kilmininkas, veikiamoji rūšis (Rimkutė, Kazlauskienė, Raškinis 2011f: 596).

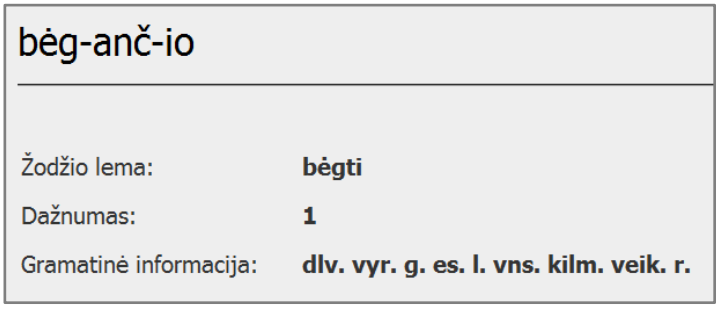
bėg-anč-io	bėgti; vksm. dlv. vyr. g. es. l. vns. kilm. veik. r.

140 pav. Informacijos apie žodį *bėgančio* pateikimas *Dažniniame lietuvių kalbos morfemikos žodyne* (Rimkutė, Kazlauskienė, Raškinis 2011f: 596)

[162] «Причастие – самостоятельная часть речи, которая обозначает признак предмета по действию, объединяет в себе свойства прилагательного и глагола» (Бабайцева, Чеснокова 2012).

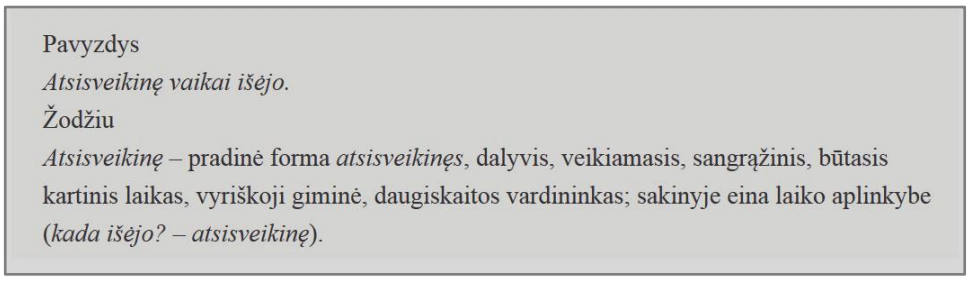
[163] “Non-finite forms of the verb” (Ambrazas 2006).

Po dvejų metų, 2013 m., VDU Kompiuterinės lingvistikos centre sukurtoje *Lietuvių kalbos morfemikos duomenų bazėje*, pateikiant informaciją apie dalyvį, jau nebenurodoma, kad tai veiksmažodis, nors duomenų bazė (50 internato nuoroda[164]) buvo parengta to paties morfemikos žodyno pagrindu. 141 pav. parodytas informacijos apie dalyvį *bėgančio* pateikimas *Lietuvių kalbos morfemikos duomenų bazėje*.



141 pav. Informacijos apie žodį *bėgančio* pateikimas *Lietuvių kalbos morfemikos duomenų bazėje* (50 internato nuoroda)

Interaktyvi lietuvių kalbos mokymosi šaltinių duomenų bazė *www.šaltiniai.info* apima ir morfologinio nagrinėjimo tvarką. Čia dalyvio pradine forma laikomas vardininkas, o ne bendratis. 142 pav. pateiktas dalyvio morfologinis nagrinėjimas sakinyje *Atsisveikinę vaikai išėjo* (94 interneto nuoroda[165]).



142 pav. Dalyvio morfologinis nagrinėjimas sakinyje (*www.šaltiniai.info*)

VII klasės vadovėlyje, kalbos dalių gramatinio nagrinėjimo lentelėje (143 pav.), dalyvis, padalyvis ir pusdalyvis išdėstomi atskirose eilutėse, kaip ir visos savarankiškos kalbos dalys (Palubinskienė, Čepaitienė 2008: 185).

Jei dalyvis būtų veiksmažodžio forma, tai jis turėtų būti lentelėje pateikiamas stulpelio pavadinime, kaip ir kitos žodžių formos, pvz., įvardžiuotinė forma būdvardžiams, sangrąžinė forma veiksmažodžiams ir kt. Taip ir dalyvis turėtų būti

[164] Prieiga internete: https://klc.vdu.lt/morfema/ [žiūrėta 2022-11-22].
[165] Prieiga internete: http://www.xn--altiniai-4wb.info/index/details/599%20 [žiūrėta 2022-11-22].

stulpelio pavadinime ir kiekvienai kalbos daliai turėtų būti nurodyta, turi ji dalyvį ar ne, taip kaip įvardžiuotinę formą vienos kalbos dalys turi, o kitos ne, pvz., veiksmažodis įvardžiuotinės formos neturi. Pateikiant dalyvį eilutės pavadinime, kur išdėstomos visos kalbos dalys, jau pripažįstama, kad dalyvis yra kalbos dalis, o ne forma.

KALBOS DALIŲ GRAMATINIS NAGRINĖJIMAS

	Kalbos dalis ar jos forma	Skyrius	Poskyris	Rūšis	Giminė	Skaičius	Linksnis	Nuosaka	Laikas	Asmuo	Laipsnis	Įvardžiuotinė forma	Sangrąžinė forma
Kaitomosios	Daiktavardis	+	–	–	+	+	+	–	–	–	–	–	+
	Būdvardis	–	–	+	+	+	+	–	–	–	+	+	–
	Įvardis	+	–	–	+	+	+	–	–	–	–	+	–
	Skaitvardis	+	+	–	+	+	+	–	–	–	–	+	–
	Veiksmažodis	–	–	–	–	+	–	+	+	+	–	–	+
	Dalyvis	–	–	+	+	+	+	–	+	–	+	+	+
	Pusdalyvis	–	–	–	+	+	–	–	–	–	–	–	+
	Padalyvis	–	–	–	–	–	–	–	+	–	–	–	+
Nekaitomosios	Prieveiksmis	+	–	–	–	–	–	–	–	–	+	–	–
	Dalelytė	–	–	–	–	–	–	–	–	–	–	–	–
	Jungtukas	–	–	–	–	–	–	–	–	–	–	–	–
	Prielinksnis	–	–	–	–	–	–	–	–	–	–	–	–
	Jaustukas	–	–	–	–	–	–	–	–	–	–	–	–
	Ištiktukas	–	–	–	–	–	–	–	–	–	–	–	–

143 pav. Kalbos dalių gramatinio nagrinėjimo lentelė VII klasės vadovėlyje (Palubinskienė, Čepaitienė 2008: 185)

2008 m. išleisto E. Palubinskienės ir G. Čepaitienės VII klasės vadovėlio lentelėje pateiktas kalbos dalių morfologinis nagrinėjimas yra aiškus ir lengvai suprantamas. Dalyvis ir veiksmažodis turi tik tris bendras morfologines kategorijas: laiką, skaičių ir sangrąžinę formą. Tačiau dalyvis turi penkis požymius, kurių neturi veiksmažodis: linksnį, giminę, rūšį, laipsnį ir įvardžiuotinę formą. Taigi, dalyvis turi daugiau skirtumų nei panašumų su veiksmažodžiu.

Dalyvio, padalyvio ir pusdalyvio atitraukimas nuo eilutės krašto gali rodyti tam tikrą priklausomybę eilute aukščiau parašytai kalbos daliai, tačiau jų išskyrimas į atskiras eilutes parodo jų savarankiškumą. Laikyti bendratį kokio nors linksnio pradine forma tiesiog nelogiška.

6.2.3. Probleminiai lietuvių kalbos dalyvio vertinimo atvejai: argumentai „už“ ir „prieš“

Tai, kad šiuo metu paplitęs dalyvio laikymas veiksmažodžio forma nėra tvirtai moksliškai pagrįstas, rodo ir kitokį požiūrį išsakantys mokslininkų teiginiai. Tiek lietuvių kalbininkų darbuose, tiek vokiečių kalbos gramatikose buvo minčių, kad dalyvis, padalyvis ir pusdalyvis yra kitos kalbos dalys nei veiksmažodis, pvz.: „Dalyvius, žyminčius daiktų ypatybes, galima ir būdvardžiais vadinti“ (Paulauskienė 2015: 201); „Padalyviai, būdami nekaitomi aplinkybės žodžiai, eina kalboje ir prieveiksmiais“ (Paulauskienė 1994: 373); „Dalyvis, kaip linksniuojamoji forma, iš tiesų yra veiksmažodinis būdvardis, kurio, kaip ir kiekvieno būdvardžio, pirmoji funkcija yra atributinė, o antroji – predikatinė, kurią gali atlikti tik vardininkas“ (Paulauskienė 2015: 356). Vokiečių kalbos gramatikoje rašoma, kad neasmenuojamos veiksmažodžio formos nesudaro specialios žodžių klasės, bet priklauso įvairioms kitoms kalbos dalims[166], pvz.: bendratis eina daiktavardžiu, dalyviai – būdvardžiu, padalyviai – prieveiksmiu ar net dalelyte[167] ir prielinksniu[168] (Helbig, Buscha 1989: 113).

Net ir aprašant analitines lietuvių kalbos veiksmažodžio formas, kurios yra bene svarbiausias argumentas laikyti dalyvį veiksmažodžio forma, nėra viskas sklandu dėl galimo dvejopo dalyvių interpretavimo. Dalyvis su asmenuojamuoju veiksmažodžiu gali būti laikomas asmenuojamosios formos sinonimu arba sudurtinio tarinio vardine dalimi. Nėra suformuluotų taisyklių, kaip juos atskirti, todėl labai dažnai tą patį sakinį galima interpretuoti dviem būdais, pvz.: *Aš jau esu išalkęs* gali reikšti *išalkau* arba *esu alkanas* (Paulauskienė 2015: 356). Toliau pripažįstama, kad „su tokia situacija vargsta iki šiol ne tik morfologijų, bet ir sintaksių autoriai“ (Paulauskienė 2015: 356).

---

[166] „Die infiniten Verbformen (Infinitiv, Partizip I, Partizip II) bilden im Deutschen keine besondere Wortklasse, sondern gehören verschiedenen anderen Wortklassen an“ (Helbig, Buscha 1989: 113).

[167] „Er wollte es *brennend* gern wissen“, „*Ausgerechnet* ihn traf ich“ (Helbig, Buscha 1989: 113).

[168] „Er wird *entsprechend* seinen Leistungen bezahlt“ (Helbig, Buscha 1989: 113).

6.2.3.1. Žodžių skirstymo į kalbos dalis kriterijai gramatikose ir mokomuosiuose leidiniuose

Internete pateikiamoje informacijoje nurodoma, kad kalbos dalis – tai kategorija, apimanti žodžius, turinčius panašias gramatines ypatybes. Žodžiai, kurie priskiriami tai pačiai kalbos daliai, turi panašius sintaksinius požymius – atlieka tas pačias funkcijas sakinyje, o morfologijos požiūriu jie turi vienodą kaitybą (88 interneto nuoroda[169]). Kalbos dalys – tai žodžių klasės, skiriamos pagal reikšmės, sintaksinių ryšių ir morfologinių požymių bendrumą (95 interneto nuoroda[170]).

XX a. *Lietuvių kalbos gramatikoje* rašoma, kad kalbos dalys – tai žodžių klasės, kurios išsiskiria tam tikrais požymiais. Jų išvardijami keturi: 1) leksinė reikšmė, 2) gramatinės kategorijos, 3) sintaksinė funkcija ir 4) žodžių darybos priemonės (Ulvydas 1965: 28). Toliau ši gramatika pripažįsta „tradicinio kalbos dalių skirstymo ribotumą, nepakankamą skirstymo kriterijų tvirtumą“ (Ulvydas 1965: 29). Kad pirmojo kriterijaus neužtenka, parodoma teiginiu apie daiktavardžio reikšmę: jis gali reikšti tiek daiktą, tiek veiksmą, tiek savybę (Ulvydas 1965: 30). Antrasis požymis – gramatinių kategorijų ir su jomis susijusių formų sistema – išryškina nenuoseklumą interpretuojant dalyvį.

Lietuvių kalbos žinyne teigiama, kad „kiekvienai kaitomajai kalbos daliai būdingos vienos bendros reikšmės pagrindu susidariusios formų sistemos, vadinamos gramatinėmis kategorijomis“ (Kniūkšta 2007: 82). Toliau pateikiami pavyzdžiai: daiktavardis turi giminės, skaičiaus ir linksnio kategorijas, asmenuojamasis veiksmažodis – asmens, nuosakos, laiko ir skaičiaus kategorijas. Taigi, asmenuojamasis veiksmažodis įvardijamas kaip kaitoma kalbos dalis, būtent taip ji vadinama – ne *veiksmažodžiu*, bet *asmenuojamuoju veiksmažodžiu*. Logiškai peršasi išvada, kad *neasmenuojamasis veiksmažodis* yra jau kita kalbos dalis, nes jo gramatinės kategorijos yra kitos. Vadinasi, būtų tikslinga dalyvį (neasmenuojamąjį veiksmažodį) ir laikyti kita kalbos dalimi.

Toliau žinyne apie dalyvius pasakyta: „Visų laikų dalyviai gali eiti pažyminiais ir tuo jie skiriasi nuo asmenuojamųjų veiksmažodžių“ (Kniūkšta 2007: 172). Be to, „lietuvių kalba pažyminiu nevartojamos dalyvio formos neturi“ (Paulauskienė 1994: 352), vadinasi, visi dalyviai gali atlikti tą sintaksinę funkciją, kurios

[169] Prieiga internete: https://en.wikipedia.org/wiki/Part_of_speech [žiūrėta 2022-11-22].
[170] Prieiga internete: https://lt.wikipedia.org/wiki/Kalbos_dalis [žiūrėta 2022-11-22].

veiksmažodžio asmenuojamos formos niekada neatlieka. Taigi, dar vienas požymis, kad dalyvis ir asmenuojamoji veiksmažodžio forma – ne ta pati kalbos dalis: skiriasi jų ir sintaksinė funkcija – trečiasis požymis skirstant žodžius kalbos dalimis (Ulvydas 1965: 28). Net ir naujausioje literatūroje autoriai, tyrinėjantys leksines analitines konstrukcijas, pripažįsta, kad junginiai, kai veiksmažodis yra dalyvio formos, pvz., *atliktas tyrimas*, labiau panašūs į atributinius nei į predikatinius (Kovalevskaitė, Rimkutė, Vilkaitė-Lozdienė 2020: 14).

Paminėtinas dar vienas teiginys, nurodantis dalyvį kaip kitą kalbos dalį nei veiksmažodis: „Specifinė dalyvio gramatinė kategorija yra rūšis“ (Kniūkšta 2007: 172); „Asmenuojamieji veiksmažodžiai rūšies morfemų neturi“ (Paulauskienė 1994: 351). Taigi, asmenuojamieji veiksmažodžiai neturi rūšies. Šis faktas buvo aptartas ir kituose darbuose. *Lietuvių kalbos dalyvių istorinėje sintaksėje* nurodoma, kad dalyvių morfologinės klasifikacijos pagrindu eina rūšies ir laiko kategorijos, kurios jį sieja su veiksmažodžiu (Ambrazas 1979: 16). Paskaitų lituanistams cikle keliamas klausimas: „Kaip su asmenuojamaisiais veiksmažodžiais gali jungti rūšies kategorija, kurios asmenuojamieji veiksmažodžiai morfologiškai nereiškia?“ (Paulauskienė 1994: 346).

Apibendrinant tai, kas pasakyta, akcentuotina, kad rūšies priskyrimas veiksmažodžio kategorijoms gramatikose nėra pakankamai pagrįstas ir įtikinantis. Jei kalbos dalis pavadinama terminu *veiksmažodis* ir neskaidoma į *asmenuojamąjį veiksmažodį* ir *neasmenuojamąjį veiksmažodį*: „Veiksmažodis yra savarankiška kalbos dalis, reiškianti veiksmą arba būseną ir turinti laiko, nuosakos, asmens, skaičiaus bei rūšies morfologines kategorijas“ (Ambrazas 1997: 282), tada ir gramatinės kategorijos turėtų būti išvardytos visos, taigi, ir tos, kurios būdingos dalyviui, – linksnis, giminė, o prie reiškiamų sąvokų turėtų būti pridėtas ir požymis (ne vien veiksmas ar būsena).

Gramatikose įrašyti sudėtinių veiksmažodžio formų apibūdinimai taip pat turi trūkumų, nes juose nepateikiama išsami informacija, todėl iš jų negalima susidaryti tikslaus bendro vaizdo apie tikrąją padėtį. Viename tokių apibrėžimų rašoma: „Lietuvių kalbos sudurtines veiksmažodžių formas sudaro savarankiškos leksinės reikšmės veiksmažodžio esamojo ar būtojo kartinio laiko veikiamieji arba neveikiamieji dalyviai su pagalbinio veiksmažodžio *būti* asmenuojamomis formomis“ (Ulvydas 1971: 144). Kitame apibūdinime teigiama: „Sudėtines laikų bei nuosakų formas sudaro veikiamieji esamojo arba būtojo kartinio laiko ir neveikiamieji esamojo arba būtojo laiko dalyviai su pagalbinio veiksmažodžio *būti* asmenuojamosiomis formomis“ (Ambrazas 1997: 346). Jau vien iš šių apibrėžimų matyti, kad

veiksmažodžio formoms sudaryti naudojami ne visi dalyvių laikai. Be to, abiejuose apibrėžimuose net neužsiminta apie tai, kad sudėtinėms formoms sudaryti naudojamas tik dalyvių vardininko linksnis. Kiti linksniai nėra veiksmažodžio formos, nes nesudaro veiksmažodžio laikų, pvz., nėra tokios sudėtinio laiko formos lietuvių kalboje: *aš esu mačiusį*. Taigi, daug dalyvio formų iš viso nenaudojamos veiksmažodžio laikams sudaryti: visi linksniai, išskyrus vardininką, būsimasis ir būtasis dažninis laikai, kurių net vardininkas nenaudojamas sudėtiniam veiksmažodžio laikui sudaryti. Ir visa tai laikoma veiksmažodžio formomis!? Tuo labiau kad net ir dalyvio laikas yra ne to paties pobūdžio kaip veiksmažodžio. Veiksmažodis reiškia sakinio predikatinį laiką (rodo santykį su kalbamuoju momentu), o dalyvio laikas yra susijęs su paties žodžio semantika – jis neturi nieko bendra su tarinio, t. y. juo einančio veiksmažodžio, laiku: *esu pavargęs*, *buvau pavargęs*, *būdavau pavargęs*, *būsiu pavargęs*. Taigi, tai dar viena prielaida manyti, kad dalyvis yra kita nei veiksmažodis kalbos dalis. Yra buvę pasiūlymų sudėtines laikų formas atriboti nuo veiksmažodžio morfologinės sistemos ir laikyti jas sintaksinėmis konstrukcijomis: „Sunku pagrįsti ir tokių konstrukcijų, kaip *yra matęs*, [...] *yra mušamas*, laikymą veiksmažodžio formomis“ (Girdenis, Žulys 1973: 208); „Kita kalbininkų grupė (Aleksas Girdenis ir Vladas Žulys, Aldona Paulauskienė, Axelis Holvoetas su Jūrate Pajėdiene ir kt.) mano, kad aptariamieji junginiai neturėtų būti laikomi kaitybinėmis veiksmažodžio formomis, nes nėra pakankamai sugramatinti. Šie kalbininkai remiasi laisva tokius junginius sudarančių žodžių tvarka (buvo vejamas: vejamas buvo), galimybe tarp jų įterpti kitus žodžius (buvo smarkiai vejamas), veikslo kategorijos priešpriešomis (buvo vejamas: pabuvo vejamas / buvo pavejamas), neigimo ypatybėmis (buvo vejamas: nebuvo vejamas / buvo nevejamas / nebuvo nevejamas)“ (Judžentis 2012: 140). Panašias mintis išsako ir kiti autoriai: „Tose kalbose, kur sudėtinga laikų sistema, [...] nekyla abejonių dėl analitinių formų priklausymo laikų paradigmai [...]. Lietuvių kalboje [...], veikiant apskritai analitinių formų neturinčiai kalbai, analitiškumas čia nyksta“ (Paulauskienė 2001: 225).

Literatūroje pateikiamas dar vienas argumentas dalyvį laikyti veiksmažodžio forma: dalyvius su veiksmažodžiais sieja ir tai, kad „veikiamieji dalyviai gali valdyti tuos pačius linksnius kaip atitinkami veiksmažodžiai“ (Paulauskienė 1994: 346). Tačiau neveikiamieji dalyviai nevaldo tų pačių linksnių kaip veiksmažodis, pvz., *skaito* ką? *skaitantis knygą* ką? – veikiamosios rūšies dalyvis valdo galininką kaip ir veiksmažodis, taičiau *skaitoma knyga* (neveikiamosios rūšies dalyvis) jau nevaldo

galininko. Taigi, dalyvius mėginama įsprausti į veiksmažodžio rėmus remiantis kriterijais, kurių kiekvienas tinka tik tam tikrai grupei dalyvių. Toks iš gabalų sudėliotas, kupinas prieštaravimų, dalyvio priskyrimas veiksmažodžiui neatrodo labai teisingas. Ta pati autorė kitame leidinyje teigia, kad dalyvių daryba, kirčiavimas, linksniavimas ir vartojimas aprašyti labai išsamiai ir abejonių nekelia, bet „teorija ir logika gerokai šlubuoja“ (Paulauskienė 2015: 354). Prieinama net prie tokios išvados, kad reikėtų keisti morfologijos sampratą: „Dalyvio laikymas veiksmažodžio forma dabar labai nedera su gerokai pasiaurinta, sukonkretinta morfologijos sąvoka“ (Paulauskienė 1994: 347). Tiktai nenurodoma, kada morfologijos sąvoka buvo platesnė ir kada ji susiaurėjo. Toliau autorė teigia: „Norint sukurti nuoseklią gramatinę dalyvio teoriją, būtina plėsti morfologijos sąvoką“ (Paulauskienė 1994: 347). Vadinasi, morfologijos mokslą siūloma priderinti prie dalyvio reikmių, t. y. reikmės laikyti jį veiksmažodžio forma. Ar ne paprasčiau būtų dalyvį laikyti savarankiška kalbos dalimi šiuolaikinės morfologijos rėmuose?

Dalyvio pateikimas mokykloms skirtose mokymo priemonėse. Painiavą galima įžvelgti ir duomenų bazės *www.šaltiniai.info* morfologinio nagrinėjimo tvarkoje: linksniuojamoms kalbos dalims ir veiksmažodžiui nurodoma kalbos dalis (atitinkamai 144 pav. ir 145 pav.). Dalyviui (146 pav.) kalbos dalies nurodyti nereikia, jį nagrinėjant turi būti pateikta veiksmažodžio forma (94 interneto nuoroda[171]).

Linksniuojamųjų kalbos dalių gramatinio nagrinėjimo tvarka yra tokia:
pagrindinė forma;
kalbos dalis;
skyrius;
rūšis (įvardžiuotinis ar paprastasis);
linksniuotė (jei turi);
laipsnis (jei turi);
giminė (jei turi);
skaičius ir linksnis;
kuo eina sakinyje.

144 pav. Linksniuojamųjų kalbos dalių gramatinio nagrinėjimo tvarka (*www.šaltiniai.info*)

[171] Prieiga internete: http://www.šaltiniai.info/index/details/599%20 [žiūrėta 2022-11-22].

Veiksmažodžio gramatinio nagrinėjimo tvarka yra tokia:
1. pagrindinės formos;
2. kalbos dalis;
3. asmenuotė;
4. nuosaka, laikas, skaičius, asmuo;
5. kuo eina sakinyje.
Pastaba. Jei veiksmažodis sangrąžinis arba beasmenis, tai nurodoma prieš asmenuotę.
Bendratis, kaip neasmenuojama forma, smulkiau nenagrinėjama.

145 pav. Veiksmažodžio gramatinio nagrinėjimo tvarka (*www.šaltiniai.info*)

Dalyvio gramatinio nagrinėjimo tvarka yra tokia:
1. pradinė forma;
2. veiksmažodžio forma;
3. rūšis;
4. laikas;
5. giminė, skaičius, linksnis;
6. kuo eina sakinyje.

146 pav. Dalyvio gramatinio nagrinėjimo tvarka (*www.šaltiniai.info*)

Daug logiškiau ir paprasčiau būtų sakyti, kad dalyvis yra kalbos dalis, nes nagrinėjimo apraše jis ir pateikiamas kaip kalbos dalis, t. y. kalbos dalies pozicijoje. Būtų aiškiau, jei visi žodžiai būtų nagrinėjami vienodai, t. y. nurodant kalbos dalį.

6.2.3.2. Dalyvio statuso pagrindimas

Kitose kalbose dalyvio laikymas veiksmažodžio forma yra labiau pagrįstas, nes analitinėse kalbose tai iš tikrųjų tik forma, o ne formų rinkinys, t. y. ne visa paradigma kaip lietuvių kalboje. Tiek anglų, tiek vokiečių kalbose dalyvis yra viena iš pagrindinių veiksmažodžio formų. Internete pateikiami duomenys, kad vokiečių kalbos veiksmažodis turi tris pagrindines formas: bendratį, būtąjį laiką ir dalyvį[172], bei nurodomi keli ir taisyklingųjų veiksmažodžių (pvz., *lernen–lernte–gelernt*), ir

[172] „Das deutsche Verb hat drei Grundformen: Infinitiv-Präteritum-Partizip II“ (88 interneto nuoroda).

netaisyklingųjų veiksmažodžių (pvz., *sprechen-sprach-gesprochen*) pavyzdžiai (96 interneto nuoroda[173]).

Anglų kalboje dalyvis taip pat priskiriamas pagrindinėms veiksmažodžio formoms. *Kembridžo žodyno* (angl. *Cambridge Dictionary*) gramatikos aprašyme teigiama, kad veiksmažodžiai turi tris pagrindines formas: pradinę (bendratį su dalelyte *to* arba be jos), būtojo laiko formą ir dalyvio (*-ed* formą), bei pateikiama vartojimo pavyzdžių[174] (97 interneto nuoroda[175]).

Lietuvių kalboje dalyvis nepriklauso pagrindinėms veiksmažodžio formoms. *Dabartinės lietuvių kalbos žodyne* rašoma: „Veiksmažodžių antraštine forma eina bendratis, po kurios pateikiamos ir kitos dvi pagrindinės formos – esamojo ir būtojo kartinio laiko trečiųjų asmenų formos“ (Keinys 1993: X). Vadinasi, lietuvių kalboje dalyvis nėra pagrindinė veiksmažodžio forma ir apskritai abejotina, ar tai yra veiksmažodžio forma.

Rusų kalbos mokyklinėje gramatikoje nurodoma, kad dalyvis yra savarankiška kalbos dalis (Бабайцева, Чеснокова 2012: 167). Akademinėje rusų kalbos gramatikoje dalyvis laikomas veiksmažodžio forma, tačiau pripažįstama, kad dalyviai turi neveiksmažodinę kaitybą – jie sudaro linksnių formas[176] (Шведова 1980: 662). Nelogiška priskirti kalbos daliai (šiuo atveju veiksmažodžiui) žodžius, kurie kaitomi jai nebūdingais linksniais.

---

[173] Prieiga internete: https://www.lingq.com/lesson/25-drei-grundformen-der-verben-468764/ [žiūrėta 2022-11-22].

[174] “Main verbs have three basic forms: base form, the past form and the -ed form (sometimes called the ‘-ed participle): base form: used as infinitive form with *to* or without *to* (*Do you want to come with us? I can’t leave now.*); past form: used for the past simple (*He opened the door and went out.*); -ed form: used after auxiliary *have* and *be* (*I’ve always wanted a piano and I was given one last week.*)” (97 interneto nuoroda).

[175] Prieiga internete: http://dictionary.cambridge.org/grammar/british-grammar/about-verbs/verbs-basic-forms [žiūrėta 2022-11-22].

[176] «Причастия обладают неглагольным словоизменением: они образуют падежные формы по адъективному склонению» (Шведова 1980: 662).

6.3. LIGIS išgauta informacija netyrinėtais lietuvių kalbos klausimais

Iki XXI a. pradžios visi klausimai, susiję su gramatika, buvo analizuojami iš žmogaus, kaip kalbos vartotojo, pozicijų, t. y. aprašomi tik tie aspektai, kurie aktualūs žmogui, ir net neužsimenama apie dalykus, kurie žmonėms yra „savaime suprantami“. Pastaruoju metu, pradėjus kalbos reiškinius kompiuterizuoti, iškyla daug problemų, kurių kalbos vartotojai paprastai net nepastebi, tačiau su jomis susiduria kompiuterinės lingvistikos specialistai, atlikdami kompiuterinio kalbos apdorojimo užduotis. Apdorojant kalbą kompiuteriu, spausdintose gramatikose ir žodynuose esanti informacija yra nepakankama, nes ten pateikiami tik bendriausieji lietuvių kalbos dėsningumai. Tai atsispindi ir kitų mokslininkų teiginiuose. Erika Rimkutė savo straipsniuose rašo: „Tačiau žmonėms skirti žodynai sunkiai pritaikomi kuriant kalbos apdorojimo programas“ (Rimkutė, Kovalevskaitė 2008b: 3); „DLKŽ [...] dažniausiai pateikti tik pamatiniai žodžiai“ (Rimkutė 2006: 38).

Kompiuterizuojant lietuvių kalbą, buvo pastebėtas morfologinis reiškinys, kurio aprašyto publikacijose nepavyko rasti. Tai paradigmose nurodytų kaitybinių formų ir galimybės jas vartoti neatitikimas. Ši problema išryškėjo tik kompiuterinio kalbos apdorojimo metu ir anksčiau kalbininkų nebuvo pastebėta. Terminu *paradigma* morfologijoje įvardijama „tam tikro žodžių kaitybos tipo galūnių sistema ar atskiro žodžio galūninių formų visuma, laikoma atitinkamo kaitybos tipo pavyzdžiu“ (Gaivenis, Keinys 1990: 143). Pavyzdžiui, veiksmažodžių esamojo laiko formos sudaromos iš esamojo laiko kamieno ir asmenų galūnių. Esamuoju laiku veiksmažodžiai asmenuojami pagal tris paradigmas, kurios atitinka tris asmenuotes (Ambrazas 1997: 337), o vieną daiktavardžių linksniuotę (*i*)*a* sudaro trys paradigmos, kurios viena nuo kitos skiriasi kai kurių linksnių formomis (Ambrazas 1999: 70). Taigi, visos kaitomos kalbos dalys turi savo paradigmas. Kol kas buvo tyrinėtos tik veiksmažodžių paradigmos. Paaiškėjo, kad ne visos žodžių formos yra galimos. Toliau aprašomi nustatyti atvejai, kai tam tikros žodžių formos negali būti vartojamos.

6.3.1. Gramatikose aptartos nevartojamos formos

Tiek vienatomėje (Ambrazas 1997), tiek tritomėje (Ulvydas 1971) gramatikose aptariami veiksmažodžio tam tikrų formų nevartojimo atvejai. Rodikliu čia gali būti laikomas ir semantinis požymis – tai vadinamieji beasmeniai veiksmažodžiai. Beasmeniai veiksmažodžiai nusako nuo veikėjo valios nepriklausančius gamtos reiškinius ar kitus stichinius procesus ir turi tik trečiojo asmens formas. Beasmenių veiksmažodžių paradigmą sudaro tiesioginės, tariamosios ir netiesioginės nuosakų 3 asmens formos, bendratis, padalyviai, kai kurių veiksmažodžių – ir neveikiamųjų dalyvių bevardė giminė (Ambrazas 1997: 315–316). Panašiai rašoma ir akademinėje gramatikoje. „Beasmeniai veiksmažodžiai žymi veiksmus ir būsenas, kurios nesiejamos su jokiu veikėju. Jie nekaitomi nei asmenimis, nei skaičiais ir savo forma sutampa su asmeninių veiksmažodžių trečiuoju asmeniu“ (Ulvydas 1971: 213). Tačiau daugumos išvardytų žodžių, kurie pateikiami kaip beasmenių veiksmažodžių pavyzdžiai, beasmeniškumas yra sąlyginis, t. y. beasmeniai jie būna tik tam tikrame kontekste, kai sakinyje nėra veiksnio. Beveik visiems gramatikoje paminėtiems beasmeniams veiksmažodžiams galima rasti tautologinį veiksnį: *lietus lyja*, *sninga sniegas* ir t. t. Be to, ir dainos žodžiais sakoma: „Ar diena dienos, ar naktelė tems“ (98 interneto nuoroda[177]). Tada ir teiginiai, kad negalimos šių žodžių liepiamosios nuosakos formos, nėra visiškai tikslūs, pvz., *Lyk, lietuti, lyk*. Net ir tokių veiksmažodžių, kuriems negalima rasti veiksnio, pvz., *reikėti*, taip pat vartojamos kai kurios liepiamosios nuosakos formos, pvz., vienaskaitos antrasis asmuo. Jis pasitaiko įvairiuose posakiuose, pvz., tekstyne: „Reikėk šitaip atsitikti“; interneto įrašuose: „Dingo. Viskas dingo. Ir reikėk tu man. Vėl ten pasirodyti.“ Panašių pavyzdžių galima rasti ir grožinėje literatūroje, pvz.: „Reikėk tu man Polionijos seseriai sykį išsižioti apie tą vokietuką“ (Aputis 1977: 32).

Toliau aptariant beasmenių veiksmažodžių vartoseną *Dabartinėje lietuvių kalbos gramatikoje* pasakyta, kad „[...] ištisas („pilnas“) paradigmas turi ne visi beasmeniai veiksmažodžiai. Pavyzdžiui, veiksmažodžių *pabaiso*, *pagailo*, *pagardo* paprastai vartojamas tik būtasis kartinis, rečiau – būtasis dažninis laikas, o kitas formas atstoja konstrukcijos su atitinkamos reikšmės būdvardžių bevarde gimine (*baisu*, *gardu*) arba

[177] Prieiga internete: https://www.musixmatch.com/es/letras/Thundertale/%C5%BDem%C4%97j-Lietuvos [žiūrėta 2020-12-21].

prieveiksmiu *gaila* ir veiksmažodžiu *darytis*, pvz.: (*darosi*) *baisu*, *pabaiso*, *pabaisdavo*, *pasidarys baisu*; (*darosi*) *gaila*, *pagailo*, *pagaildavo* (*pasidarydavo gaila*), *pasidarys gaila*“ (Ambrazas 1997: 316). Žinoma, galima konstrukcija ir *darėsi / pasidarė gaila*. Veiksmažodis *pagailo* ir konstrukcija *darėsi / pasidarė gaila* yra sinonimiškai vartojamos formos, turinčios labai panašią ar net tą pačią reikšmę. Jei paradigmoje trūkstamas, t. y. kalboje nevartojamas, tam tikro žodžio formas gali pakeisti kitų žodžių junginiai, vadinasi, tos trūkstamos formos turi prasmę (reikšmę). Šiame darbe stengiamasi aprašyti atvejus, kai tam tikrų žodžių atskiros formos yra negalimos, t. y. žodžiai, pavartoti ta forma, neturi prasmės.

Reikia paminėti, kad minčių apie dalyvio formų nevartojimą išsakyta ir J. Jablonskio 1925 m. išleistame *Rygiškių Jono Lietuvių kalbos vadovėlyje*. Jis rašė: „Apie dalyvius „būsimas, siųsimas, vesimas“ [...] čia netenka kalbėti: jie retas tėra kalbos dalykas, jie ne visur ir ne iš visų veiksmažodžių padaromi“ (Jablonskis 1925: 61). Tačiau čia aptariamas atvejis, kai nėra visų dalyvio formų su tam tikra šaknimi.

LIGIS iškilo problema dėl dalies kaitybinių formų nebuvimo, kai kitos to paties žodžio formos yra vartojamos. O šiuo klausimu jokių publikacijų nepavyko rasti – nei Lietuvos, nei kitų šalių autorių.

6.3.2. LIGIS atlikti darbai ir iškilusios problemos

Kol kas yra sukurtas tik bandomasis LIGIS pavyzdys, apimantis šaknies *bėg-* žodžius. Buvo pasirinkta tokia darbo metodika: iš pradžių, paėmus vieną šaknį, siekta sudaryti visos sistemos modelį, t. y., panaudojant vieną šaknį, pateikti visas sistemoje numatytas informacijos rūšis, o paskui, plečiant sistemą, įtraukti ir kitų šaknų žodžius. Šaknies *bėg-* pasirinkimą lėmė jos paprastumas: niekada neatsiranda intarpas, nėra vidinės fleksijos (nevyksta priebalsių ar balsių kaita šaknyje). Siekiant apimti visas galimas visų šios šaknies žodžių formas buvo dirbtinai kompiuteriu generuojami visi vediniai su veiksmažodžiams naudojamais priešdėliais ir priesagomis. Tada kalbininkų buvo peržiūrėtos lemos ir atmesti lietuvių kalboje neegzistuojantys žodžiai. Tekstyne ieškota dūrinių su šia šaknimi. Taip gautas lemų sąrašas, kurį sudaro visos kalbos dalys, galinčios turėti *bėg-* savo šaknyje. Tai – daiktavardžiai, pvz.: *bėgikas*, *bėgtakis;* būdvardžiai, pvz.: *bėglus*, *vienbėgis (geležinkelis)*; prieveiksmiai, pvz.: *pabėgom*, *bėgčia*, *probėgšmais* ir kt. Naudojant lietuvių kalbos morfologinės sintezės programinę įrangą (Zinkevičius 2000), buvo generuojamos visos atrinktų žodžių kaitybinės formos. Iš viso duomenų bazėje pateikiami 28 322 įrašai, t. y. tiek žodžių, turinčių šaknį *bėg-*,

rasta lietuvių kalboje. Iš jų buvo: 1 254 daiktavardžiai, 330 būdvardžių, 12 prieveiksmių, 5 913 veiksmažodžių, 14 310 veikiamosios rūšies dalyvių, 3 746 neveikiamosios rūšies dalyviai, 1 772 reikiamybės dalyviai, 500 pusdalyvių, 485 padalyviai. Kol kas negalima daryti jokių apibendrinimų dėl kitų žodžių, nes atskirų šaknų produktyvumas gali labai skirtis.

Tačiau net ir atrinkus žodžių lemas, t. y. peržiūrėjus pradines formas ir išbraukus visus lietuvių kalboje neegzistuojančius morfemų rinkinius, paaiškėjo, jog to yra per maža, norint pasiekti, kad į sistemą nepakliūtų lietuvių kalboje nevartojami žodžiai. Tikrinant visas susintezuotas kaitybines formas buvo pastebėta keletas atvejų, kai to paties žodžio dalis paradigmoje esančių formų yra vartojama, o dalis – ne. Ir tą nulemia žodžio semantika, t. y. jo reikšmė. Tiksliau pasakius, taip nutinka dėl leksinės žodžio reikšmės nederėjimo su jo gramatine reikšme arba dėl jų loginio prieštaringumo.

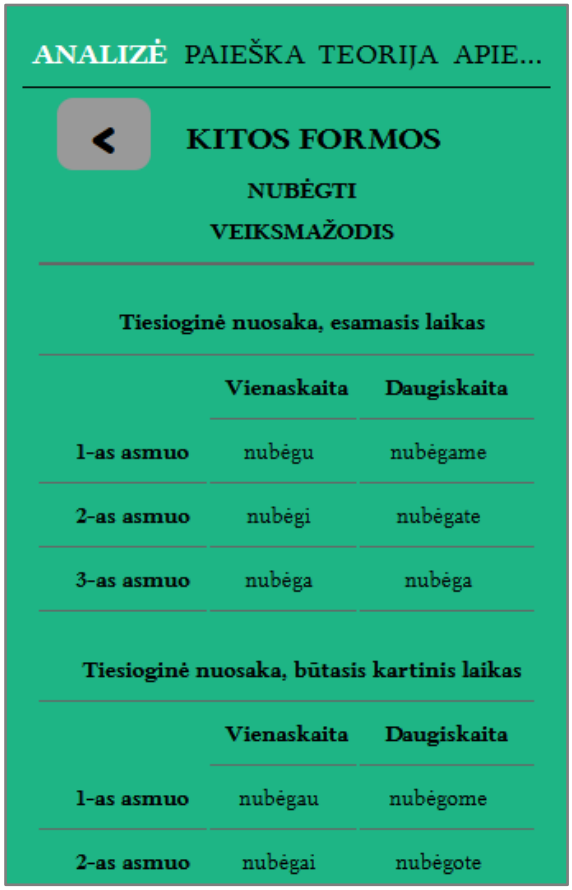
Galima būtų išskirti tris pagrindinius nevartojamų formų atvejus: a) veiksmažodžio vienaskaitos formų nebuvimą, b) veiksmažodžio būtojo kartinio laiko nebuvimą ir c) neveikiamųjų dalyvių tik bevardės giminės vartojimą. Plačiau apie tai žr. tolesniuose poskyriuose.

6.3.3. Nevartojamos veiksmažodžių formos

Atliekant VDU morfologiškai anotuoto tekstyno tyrimus, pastebėta, kad „[...] lietuvių kalba yra sudėtinga fleksinė kalba, pasižyminti gramatinių formų įvairove, tačiau realiai vartojama tik labai nedidelė tų formų dalis. Labai skiriasi gramatikose pateikiamos kalbos dalių teorinės kaitybinės paradigmos ir iš tikrųjų vartojamos formos“ (Rimkutė 2006: 39).

LIGIS viena iš plačiajai visuomenei skirtos informacijos rūšių yra vartotojo įvesto žodžio visų formų pateikimas atskirame skirtuke. Žodžio *nubėgti* paradigmos fragmentas parodytas 147 pav.

Kompiuteriu generuojant visas teoriškai galimas kaitybines formas, buvo pastebėta, kad kai kurios jų netikroviškos, ir tai nulemia konkretaus žodžio semantika.



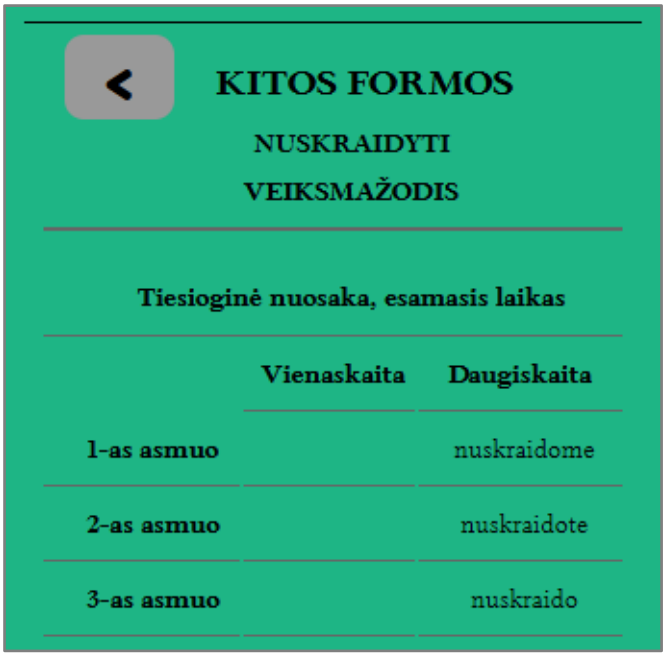
147 pav. Skirtuke KITOS FORMOS žodžiui *nubėgti* pateikiamos informacijos fragmentas

6.3.3.1. Vienaskaitos nebuvimas

Sudarant veiksmažodžio visų laikų ir nuosakų kaitybines formas paaiškėjo, kad kai kurių veiksmažodžių, ypač tų, kurie reiškia grupinį veiksmą, vienaskaitos formos yra nelogiškos, neperteikia semantinio turinio. 147 pav. pateikto žodžio *nubėgti* visos kaitybinės formos yra galimos, tačiau kito žodžio – *nuskraidyti* (turinčio tokį patį priešdėlį *nu-*) – galima tik daugiskaita, pvz., *Papūtus vėjui pro atvirą langą nuo stalo* ***nuskraidė*** *keli jos prirašyti lapai*. Šio žodžio vienaskaitos formos yra netikroviškos ir lietuviams nesuprantama, ką galėtų reikšti žodis, pvz., **nuskraidžiau*. Nekyla problemų dėl daugiskaitos pirmojo asmens, pvz., *Nuo smūgio mes visi* ***nuskraidėme*** *nuo vežimo ant žemės*. Trečiojo asmens forma vartojama tik su daugiskaitos veiksniu (sakinių pavyzdžiai paimti iš interneto): *Triūsas nuėjo perniek, nes nuo stiklainių* ***nuskraidė*** *tvirtai užsukti dangteliai. Dalis stogo liko kieme, kada* ***nuskraidė*** *kitos dalys, nežinau. Stalas parvirto, aš taip pat,* ***nuskraidė*** *tortas ir dovanos, gėlės ir lėkštės*. Internete rastas tik vienas žodžio *nuskraidžiau* pavartojimo atvejis: *Į Einthoveną* ***nuskraidžiau*** *pinigus ir*

ten juos atidaviau, o iš jo skridau į Londoną (99 interneto nuoroda[178]). Tačiau perskaičius šį sakinį tenka pamąstyti, ką galėtų reikšti *nuskraidyti pinigus*, ir atrodo, kad čia labiau tiktų žodis *nuskraidinau*. Kitas žodžio, neturinčio vienaskaitos formų, pavyzdys būtų veiksmažodis *susispiesti*. *Lietuvių kalbos žodyno* elektroniniame variante (100 interneto nuoroda[179]) pateikiami tokie jo pavartojimo pavyzdžiai: *Ei, jūs, vaikai, ko čia* ***susispietėt****? Krūvelėn* ***susispietusios*** *čionai pat stovėjo moterys. Per pirmąjį pasaulinį karą Peterburge* ***susispietė*** *nemažai lietuvių. Čia vėl* ***susispiečia*** *varnėnai į vieną juodą dėmelę. Vasarą visi darbai* ***susispiečia*** *į vieną vietą. Širdy* ***susispietė*** *jausmai kiti. Mokytojo akys* ***susispietė*** *į vieną mokinį* (Naktinienė 2017). Tačiau forma **susispiečiau* yra negalima. Dar vienas žodžio, neturinčio vienaskaitos formų, pavyzdys būtų veiksmažodis *išsiskirstyti*: *Baigę mokslo metus išsiskirstėme neramūs.* ***Išsiskirstėme*** *kaip geri draugai.* [...] *mes, buvę mokiniai, irgi* ***išsiskirstėme*** *po visą Lietuvą* [...] (DLKT). Tačiau jo forma **išsiskirsčiau* yra negalima, nes šis žodis neturi prasmės.

LIGIS ši problema sprendžiama taip: skirtuke KITOS FORMOS žodžiams, kurių vienaskaitos formos yra negalimos, pateikiama tik daugiskaita. 148 pav. parodytas žodžio *nuskraidyti* paradigmos fragmentas.



148 pav. Skirtuke KITOS FORMOS žodžiui *nuskraidyti* pateikiamos informacijos fragmentas

[178] Prieiga internete: https://laisvaslaikrastis.lt/kaip-teisejas-darius-kantaravicius-pridengineja-narkodilerius/ [žiūrėta 2019-10-10].

[179] Prieiga internete: http://www.lkz.lt/?zodis=spiesti&id=24168780000 [žiūrėta 2022-11-22].

6.3.3.2. Būtojo kartinio laiko nebuvimas

Dar vienas atvejis, kai žodžiai turi negalimų formų, taip pat susijęs su semantika. Veiksmažodžiai su priešdėliu *tebe-* reiškia tęstinį veiksmą[180] (Arkadiev 2010: 21), todėl įvykio veikslo veiksmažodžių su šiuo priešdėliu negali būti. Aprašant lietuvių kalbos gramatiką nurodoma, kad „veiksmažodžio veikslo priešpriešų forma yra nereguliari“ ir „veikslo atžvilgiu gali būti priešinamos skirtingų laikų formos – esamasis laikas turi eigos veikslo reikšmę, o būtasis ir būsimasis – įvykio [...], plg.: *Ateina vakaras nykus* (Nėris) ir *Atėjo ir lauktoji saulės užtemimo valanda* (Vienuolis)“ (Judžentis 2012: 157–158). Tačiau neatsižvelgta į tai, kad kai kurie priešdėliai gali pakeisti veikslą. Remiantis citata, žodis *ateis* yra įvykio veikslo, tačiau *tebeateis* jau rodo veiksmo tęstinumą ir yra priskiriamas eigos veikslui, lygai taip pat, kaip ir esamojo bei būtojo dažninio laiko formos *tebeateina*, *tebeateidavo*. Įvykio veikslo veiksmažodžių būtojo kartinio laiko formos su šiuo priešdėliu yra netikroviškos, pvz., trečiojo asmens forma **tebeatėjo*. LIGIS tokiems veiksmažodžiams nepateikiamos būtojo kartinio laiko formos.

Pastebėtas bendras polinkis, kad kryptį rodantys slinkties veiksmažodžiai (o kai kurie ir ne slinkties) su priešdėliais *at-*, *nu-*, prisijungę dar ir *tebe-*, negali turėti būtojo kartinio laiko formų, pvz.: *atvažiuoti*, *nuvažiuoti*, *atbėgti*, *nubėgti*, *atskristi*, *nuskristi*, *nudažyti* ir kt. Žodžiai *tebeatvažiuoja*, *tebeatvažiuodavo*, *tebeatvažiuos* galimi lietuvių kalboje, tačiau **tebeatvažiavo*, **tebenudažė* jau yra beprasmiai. Sakinys *Kai aš grįžau, jis dar tebedažė lentyną* yra suprantamas ir geras, nes *dažė* yra eigos veikslo veiksmažodis. Įvykio veikslas būtų *nudažė*, o **tebenudažė* jau skamba neįprastai, reikia įsitempus mąstyti, ką tai galėtų reikšti, ir kažin ar kokia nors šio žodžio reikšmė ateitų į galvą. *Tebe-* ir įvykio veikslas vienas kitą eliminuoja.

6.3.3.3. Vartojama tik bevardė giminė

Daugelio dalyvių vartojama tik bevardė giminė. Neveikiamosios rūšies dalyvių, padarytų iš intranzityvinių veiksmažodžių, galima tik bevardės giminės forma ir tevartojamos tik esamojo ir būtojo laiko formos. Tai matyti ir iš statistinių LIGIS duomenų. Veikiamosios rūšies dalyvių su šaknimi *bėg-* yra 14 310, o neveikiamosios

[180] “[...] the combination of *te-* and *be-* [...] this sequence can only be interpreted as a continuative form” (Arkadiev 2010: 21).

rūšies dalyvių – tik 1 772. Taigi, labai didelė dalis neveikiamosios rūšies dalyvių, t. y. tie, kurie padaryti iš intranzityvinių veiksmažodžių, neturi moteriškosios ir vyriškosios giminės formų bei linksnių, pvz., žodžių junginiai *bėgamas ruožas*, *prabėgta atkarpa*, *nubėgti kilometrai* yra suprantami taip pat, kaip ir *vaikų jau buvo išsibėgiota kas kur*. Tačiau žodis **išsibėgiotas* jau yra negalimas. Iš viso LIGIS duomenų bazė apima 236 neveikiamosios rūšies dalyvių lemas, iš jų 120 gali turėti tik bevardės giminės formas.

Užuominų apie dalyvių, padarytų iš intranzityvinių veiksmažodžių, vardininko formos nebuvimą, galima pastebėti ir *Dabartinės lietuvių kalbos gramatikoje*. Aprašant sudėtines laikų formas, pasakyta: „Objekto linksnio nevaldantys veiksmažodžiai turi sudėtines neveikiamąsias formas tik su bevardės giminės dalyviais (yra / buvo [...] vaikščiojama, vaikščiota)“ (Ambrazas 1997: 320).

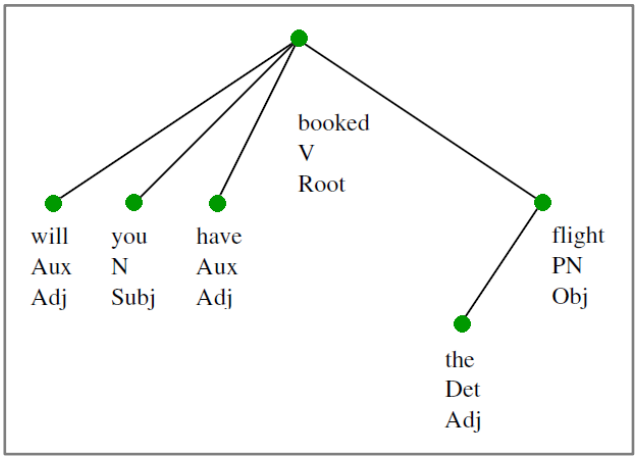
6.4. LIGIS perspektyva – sintaksės dalis

Šiuo metu yra parengtas bandomasis pavyzdys tik vienai gramatikos daliai – morfologijai. Sintaksinę dalį planuojama kurti ateityje. Ruošiant informaciją plačiajai visuomenei buvo stengiamasi kuo aiškiau ir suprantamiau pateikti žodžių morfologinius požymius. Tokia pati nuostata išlieka ir sintaksei. Svarbiausia, sudarant sakinio sintaksinę struktūrą, bus siekiama kuo tiksliau perteikti lietuvių kalbos specifinius bruožus ir iš visų kitoms kalboms taikomų metodikų bus atsirenkamos tinkamiausios, nė vienos iš jų neperimant aklai, kaip vienintelės galimos ir neklystančios.

Sakinys iš pradžių išskaidomas į veiksnio ir tarinio grupes bei pabaigos ženklą. „Subjekto ir predikato dichotomija turi ilgą tradiciją, siekiančią senovės gramatikas. XX amžiaus lingvistikoje šios koncepcijos iš dalies atsisakyta, vis dėlto ji [...] iki šiol dar gyvuoja“ (Holvoet 2009: 149). Taigi, ją pasirinkome ir LIGIS, nes ir kituose lietuvių kalbininkų darbuose teigiama: „Visą sakinį pirmiausia galima struktūriškai skaidyti į veiksnio ir tarinio grupes, kurių vienos centre – veiksnys, o kitos – tarinys. Pagal svarbą ir struktūrinį vaidmenį sakinyje yra pagrindo veiksnį ir tarinį laikyti pirmojo rango sakinio dalimis, o visos kitos sakinio dalys būtų žemesnių rangų“ (Labutis 2002: 202).

Verbocentrinė sakinio struktūros forma nėra pati tinkamiausia lietuviškam sakiniui pavaizduoti. Ji atrodo nelogiška net ir angliškame sakinyje, kai veiksnys yra

tame pačiame lygmenyje kartu su nesavarankiškais žodžiais, pvz., 149 pav. pateiktas *Pensilvanijos sintaksiškai anotuoto tekstyno* sakinys, kurio struktūroje matyti, kad žodis *you*, t. y. sakinio veiksnys, yra tame pačiame lygmenyje, kaip ir pagalbinis veiksmažodis *will* (Rambow ir kt. 2000: 2), kuris naudojamas tik laiko formai sudaryti, panašiai kaip lietuvių kalbos priesaga *-dav-*, su kuria padaromas būtasis dažninis laikas.



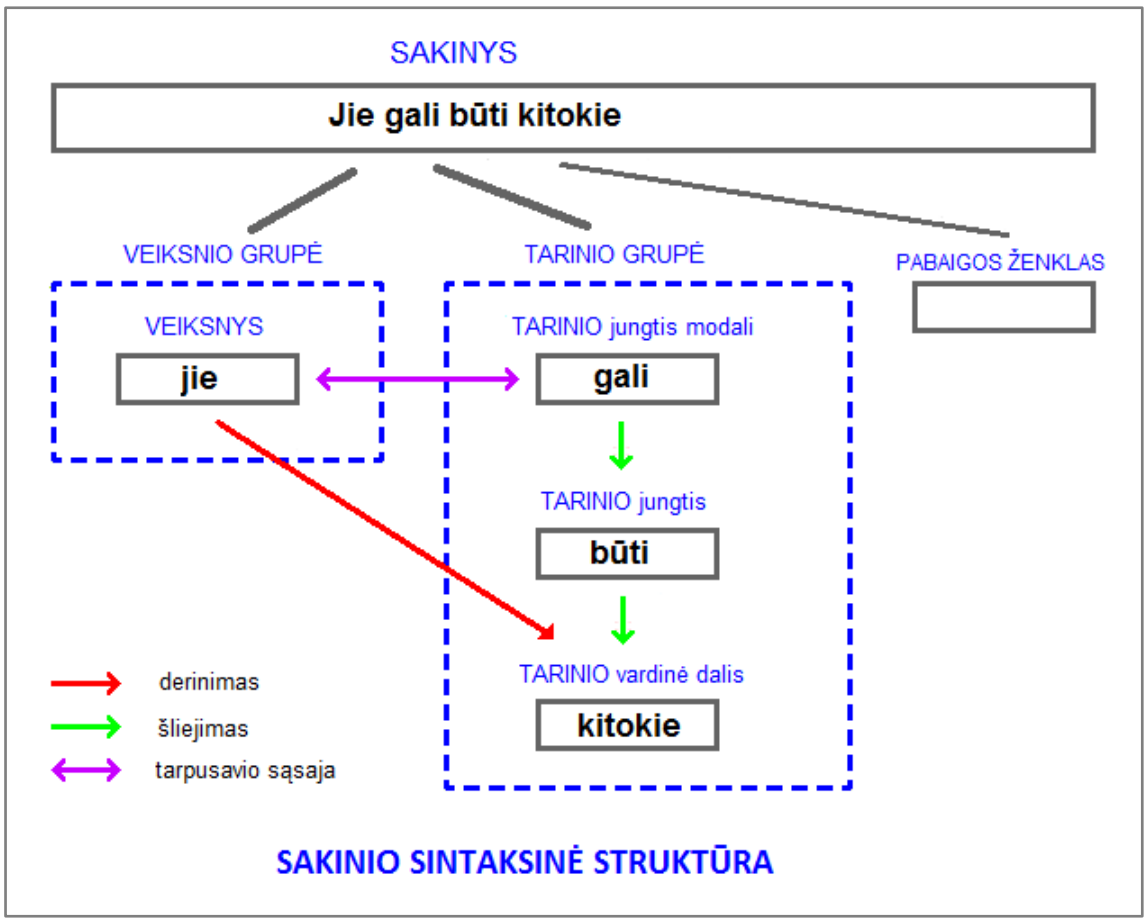
149 pav. Sakinio *will you have booked the flight* analizė (Rambow ir kt. 2000: 2)

Lietuviškam sakiniui parenkant sintaksinės struktūros pavaizdavimo būdą, turėtų būti įvertintos specifinės lietuvių kalbos ypatybės ir iš visų pasaulyje sukurtų metodikų paimama tai, kas mums tikrai tinka. Taip elgiasi ir kitų šalių kalbininkai, pvz., vokiečių mokslininkai Ulrichas Engelis (Ulrich Engel), Hansas Verneris Eromsas (Hans Werner Eroms), Hansas Jurgenas Heringeris (Hans Jürgen Heringer) savo darbuose įvertina tik vokiečių kalbos gramatikai būdingas savybes[181] (Agel 2000: 92).

Pasirinkus frazių gramatikoje naudojamą sakinio skaidymą į veiksnio ir tarinio grupes, šios gramatikos taikymas tuo ir baigiamas. Lietuvių kalbai, kuriai būdinga laisva žodžių tvarka (kartais žodžių tvarka pakeičiama stilistiniais sumetimais, siekiant išvengti monotoniškumo), netikslinga toliau taikyti frazių gramatikos, nes labai daug sakinių būtų neprojektyvūs. Pasitaiko atvejų, kai tarp papildinio ir jo pažyminio įsiterpia veiksnys (kartais net su kreipiniu), pvz., ***Geras*** *gi tu, Simanai,* ***akis*** *turi, kad galėjai ją [žvaigždę] pamatyti* (Ambrazas 1997: 432), ar tarinys, pvz., *Iš mažos kibirkšties* ***didis*** *kyla* ***gaisras*** (Ambrazas 1997: 655), ar net abu kartu, t. y. ir veiksnys, ir tarinys, pvz., ***Liūdną*** *jis mums pranešė* ***žinią****: Mykolas sėdi kalėjime.*

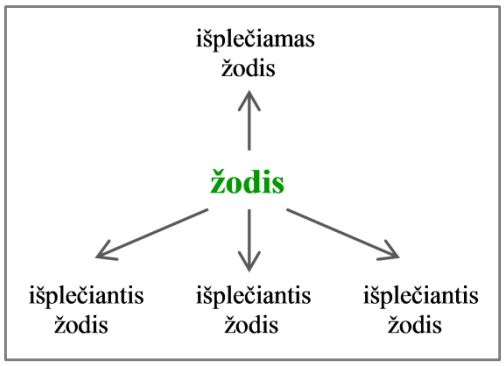
[181] „[...] ziehen sie in ihren Arbeiten [...] nur grammatische Eigenschaften des Deutschen heran“ (Agel 2000: 92).

Buvo atlikti rusų kalbos (turinčios laisvą žodžių tvarką) tyrimai, siekiant nustatyti, kokiu atstumu sakinyje gali būti tiesioginiu sintaksiniu ryšiu susiję žodžiai, nes sintaksinės analizės metu daug problemų iškyla nustatant prielinksnio ryšių zoną. Ukrainos mokslų akademijoje, Kijeve, atliktų tyrimų rezultatas: didžiausias užfiksuotas atstumas tarp priklausomų žodžių – 28, t. y. tarp dviejų tiesioginiu sintaksiniu ryšiu susietų žodžių buvo įsiterpę 28 kiti žodžiai (Грязнухина 1999: 136). Taigi, tokiais atvejais mažai tikėtina, kad sakinio sintaksinę struktūrą frazių gramatikoje pavyks pavaizduoti be linijų susikirtimo. Lietuvių kalbos rezultatai gali būti panašūs. Todėl, išskaidžius sakinį į veiksnio ir tarinio grupes, toliau sintaksinė struktūra vaizduojama vadovaujantis priklausomybių gramatikos principu. Iš tekstyno paimto sakinio *Jie gali būti kitokie* analizė pateikta 150 pav. Sakinio pabaigoje nėra jokio ženklo, nes tai straipsnelio pavadinimas.



150 pav. Sakinio *Jie gali būti kitokie* sintaksinė struktūra

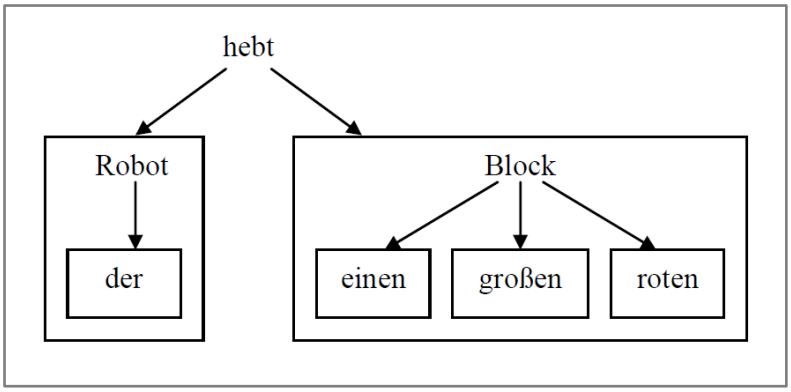
Apibendrinta priklausomybių gramatikos viršūnė pavaizduota 151 pav. Sintaksinės analizės užduotis – surasti kiekvienam žodžiui visus nuo jo priklausomus žodžius ir nustatyti, nuo kurio žodžio jis pats priklauso (Hellwig 2002: 17).



151 pav. Apibendrinta medžio viršūnė priklausomybių gramatikoje (parengta pagal Hellwig 2002: 17)

Vokiečių mokslininkas Peteris Helvigas (Peter Hellwig) pasiūlė sudaryti priklausomybių medį pagal formalų aprašą DUG (*Dependency Unification Grammar*). Šis aprašas remiasi idėja, kad sakinyje žodį išplečia ne kuris nors vienas atskiras žodis, o visa žodžių grupė, esanti medyje žemiau jo. Tą grupę sudaro vienas tiesioginiu sintaksiniu ryšiu susietas žodis su visais jį patį išplečiančiais žodžiais.

Sakinį galima įsivaizduoti kaip sudarytą iš dėžučių, ir kiekvieną žodį išplečia visas dėžučių blokas. Rodyklės čia rodo sintaksinį ryšį ne tarp dviejų žodžių, bet tarp žodžio ir jį išplečiančios žodžių grupės (dėžučių bloko). Mokslininkas pateikė sakinio *Der Robot hebt einen großen roten Block* sintaksinių ryšių schemą (152 pav.) pagal DUG aprašą (Hellwig 2003: 13).

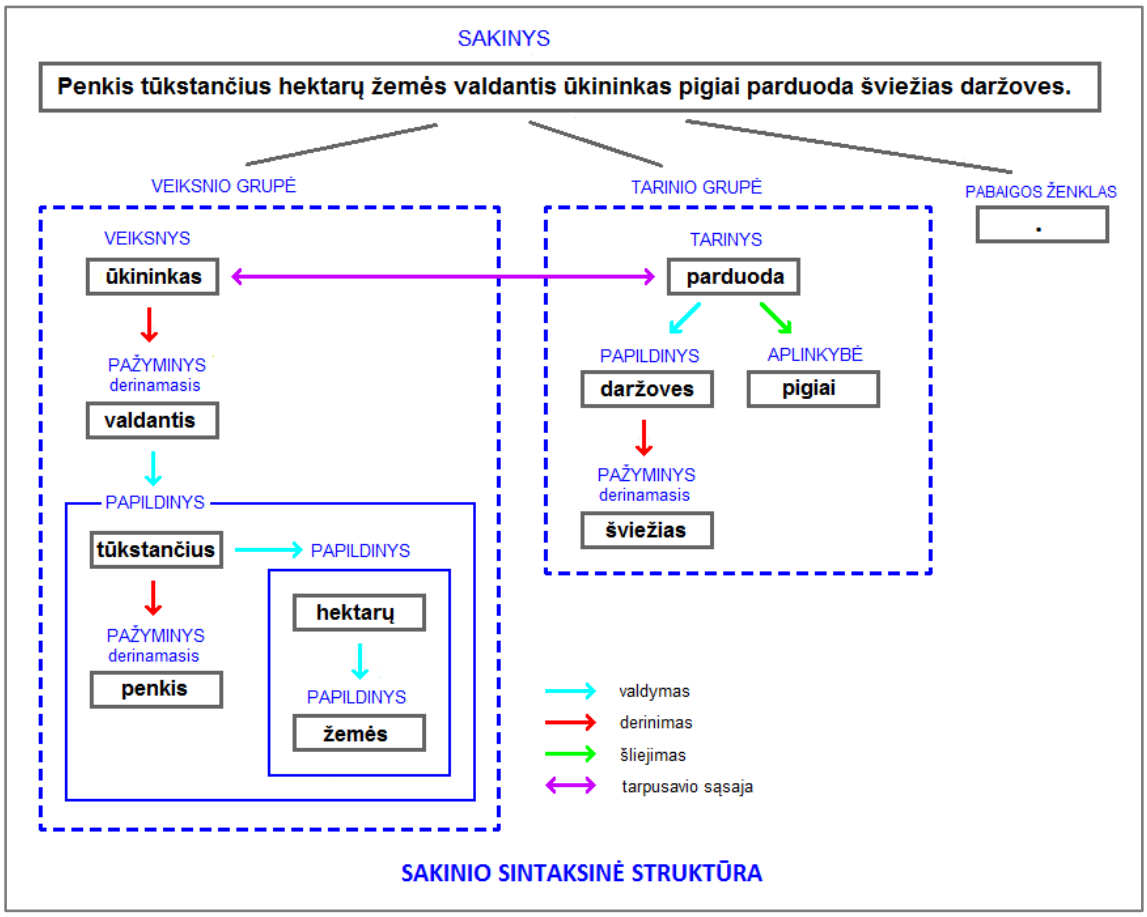


152 pav. Sakinio, sudaryto iš dėžučių, pavyzdys (Hellwig 2003: 13)

Tokia metodika, šiek tiek ją modifikavus, labai tinka kai kuriems lietuvių kalbos sakiniams pavaizduoti, ypač tais atvejais, kai naudojantis tradicine priklausomybių gramatika negalima parodyti visos sintaksinės informacijos, esančios lietuviškame sakinyje, pvz., kai pavartoti neskaidomi žodžių junginiai, frazeologizmai ir kt.

Neskaidomi žodžių junginiai – tai žodžių grupės, kurios tik kartu gali išplėsti kitą žodį, o ją išplečiantis žodis gali pažymėti tik visą grupę. Kitų sakinio žodžių ryšys su vienu iš neskaidomo junginio dėmenų neturi prasmės.

Sintaksinėje struktūroje neskaidomi junginiai traktuojami kaip vienas leksinis vienetas, todėl sudedami į vieną bloką, parodant jo vidinius sintaksinius ryšius. Kaip pavyzdį galima pateikti taip pat iš tekstyno paimto sakinio *Penkis tūkstančius hektarų žemės valdantis ūkininkas pigiai parduoda šviežias daržoves* analizės schemą (153 pav.).

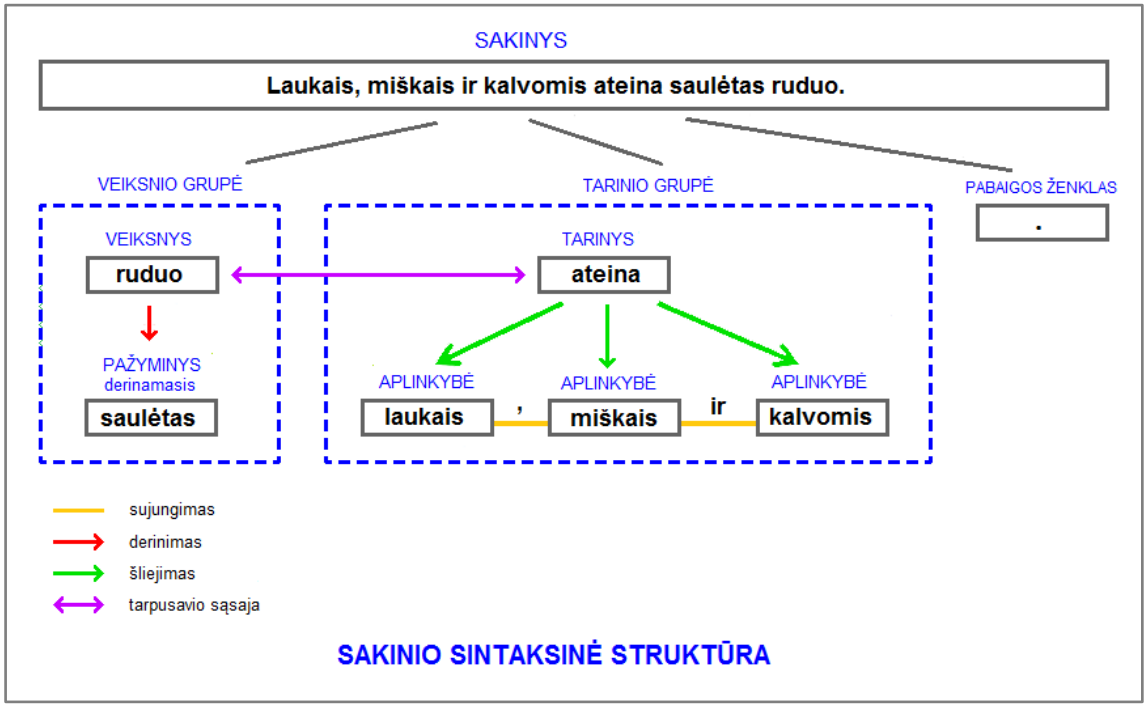


153 pav. Sakinio su neskaidomu žodžių junginiu pavyzdys

Šio sakinio dalyvis *valdantis* reikalauja papildinio, išreikšto galininko linksniu. Artimiausioje jo aplinkoje esantis galininkas yra žodis *tūkstančius*, kuris logiškai (semantiškai) nėra papildinys. Iš tikrųjų ūkininkas valdo ne skaičių, o žemę. Todėl tik visas žodžių junginys *penkis tūkstančius hektarų žemės* gali būti papildinys. Papildinio, kurį sudaro neskaidomas žodžių junginys, viduje parodyti jo dėmenų sintaksiniai ryšiai. Žodis *penkis* derinamas su *tūkstančius*, t. y. jo linksnis nulemia skaitvardžio *penki* linksnį. Jei skaitvardis *tūkstantis* būtų kito linksnio, pvz., naudininko, tada ir *penki* pasidarytų naudininko linksnio (*Penkiems tūkstančiams hektarų buvo sunaudota*

30 tonų trąšų). Skaitvardis *tūkstantis* valdo kilmininko linksnį, t. y. reikalauja jo kaip papildinio, todėl žodis *hektarų* siejamas su skaitvardžiu valdymą žyminčia rodykle. Kitas kiekio žodis – *hektarų* – taip pat reikalauja kilmininko kaip papildinio, todėl ir žodis *žemės* siejamas su žodžiu *hektarų* valdymą žyminčia rodykle.

Naudojant verbocentrinę sakinio sintaksinės struktūros formą, kuri yra patogi apdorojant kalbą kompiuteriu, sudėtinių sujungiamųjų sakinių atveju negalima pavaizduoti ryšių tarp žodžių tokių, kokie jie iš tikrųjų yra. 95 pav. pateiktoje sakinio *Mokytojas įėjo ir vaikai atsistojo* sintaksinėje struktūroje sujungiamasis jungtukas *ir*, siejantis abu sakinius, parodytas kaip priklausomas nuo antrojo sakinio *vaikai atsistojo* tarinio *atsistojo* ir turintis ryšį tik su juo vienu, nors šiaip jungtuko *ir* paskirtis yra sieti du objektus. Kad būtų išvengta panašaus tipo neatitikimų, LIGIS plačiajai visuomenei skirtoje dalyje sujungiamasis ryšys vaizduojamas geltona linija be rodyklių, nes šiuo ryšiu susieti žodžiai niekaip neveikia vienas kito. 154 pav. parodyta sakinio *Laukais, miškais ir kalvomis ateina saulėtas ruduo*, turinčio vienarūšių sakinio dalių (aplinkybių), sintaksinė struktūra. Savarankiškos reikšmės neturintys žodžiai (jungtukai, prielinksniai ir kt.) pateikiami kaip grafo lankų žymės, šiuo atveju *ir* ir kablelis.



154 pav. Sakinio *Laukais, miškais ir kalvomis ateina saulėtas ruduo* sintaksinė struktūra

6.5. Skyriaus išvados

LIGIS sukūrimo tikslas – parengti lietuvių kalbos gramatikos dokumentaciją, t. y. labai didelio tikslumo ir patikimumo gramatinius duomenis apie lietuvių kalbą.

Siekiama sukurti universalią *Lietuvių kalbos gramatikos informacinę sistemą*, kad ja galėtų naudotis tiek žmonės, tiek kompiuteriai, t. y. kad ji populiariai pateiktų informaciją plačiajai visuomenei ir kad būtų tinkama kompiuteriniam lietuvių kalbos apdorojimui.

Plačiajai visuomenei skirtoje morfologinėje dalyje populiariai pateikiami trijų tipų duomenys apie žodį: morfologiniai, morfeminiai ir darybiniai. Svarbu, kad LIGIS pirmą kartą viešai pateikiama informacija apie morfemos tipą ir jos charakteristikas. Taip pat galima pamatyti kiekvieno žodžio visas kaitybines formas.

Morfologinės žymos, skirtos kalbai apdoroti kompiuteriu, buvo sudaromos savos, nes nė vienas iš jau sukurtų ir naudojamų lietuvių kalbos žymų rinkinių netenkino visų LIGIS poreikių.

Kuriant LIGIS paaiškėjo keli nauji faktai apie lietuvių kalbą ir jos gramatiką, į kuriuos anksčiau nebuvo atkreiptas dėmesys, nes žmogui, kaip kalbos vartotojui, tie klausimai neaktualūs. Kompiuterizuojant lietuvių kalbos gramatiką pastebėta, kad kai kurie lietuvių kalbos žodžiai negali įgyti visų teorinės paradigmos formų ir šį reiškinį nulemia semantika: taip nutinka dėl leksinės reikšmės nederėjimo su gramatine reikšme arba dėl jų loginio prieštaringumo. Grupinį veiksmą reiškiančių žodžių vienaskaita yra negalima. Įvykio veikslo veiksmažodžių vediniai su priešdėliu *tebe-* negali turėti būtojo kartinio laiko. Neveikiamosios rūšies dalyviai, padaryti iš intranzityvinių veiksmažodžių, gali turėti tik bevardės giminės formą.

Taigi, apdorojant kalbas kompiuteriu, galima pažvelgti į gramatiką iš kitos pusės ir pastebėti žmogui nekrintančią į akis informaciją.

APIBENDRINAMOSIOS IŠVADOS

Šiuo metu kompiuterizuojant tautų kalbas, dažniausiai naudojami statistiniai ir neuroninių tinklų metodai. Jiems būdinga tai, kad labai greitai sukuriamos galingos sistemos, gebančios apdoroti nepaprastai dideles žodžių apimtis, bet tik su viena sąlyga – turi būti toleruojamas tam tikras kiekis klaidų. Ir jau atvirai pripažįstama, kad 100 proc. tikslumas gali būti niekada ir nepasiektas. Lietuvių–anglų ir anglų–lietuvių kalbų automatinis vertimas *Tilde*, naudojantis statistinius metodus (2017), pateikia geresnius rezultatus negu neuroniniais tinklais pagrįstas tų pačių kalbų vertimas (2019).

Atliekant automatinę morfologinę analizę, naudojami du pagrindiniai metodai: taisyklėmis pagrįstas ir statistinis. Norint sukurti taisyklėmis pagrįstu metodu veikiantį analizatorių, reikia įdėti daug labai aukštos kvalifikacijos specialistų darbo, bet jam nereikia parengti jokių papildomų išteklių, tokių kaip ranka anotuoti tekstynai. Šiuo metodu veikianti programinė įranga pateikia labai tikslią analizę, bet lieka daugiareikšmiškumas. Statistiniais metodais dirbančioms sistemoms reikia palyginti nedidelio kiekio žmogaus anotuotų sakinių, ir jos gali anotuoti labai didelės apimties tekstynus. Tačiau rezultatai gaunami su tam tikra klaidos tikimybe, t. y. absoliutaus tikslumo čia nepasiekiama. Statistiniai metodai gana gerai tinka skaidant žodžius morfemomis. Ypač jie vertingi tais atvejais, kai reikia sudaryti mažai tyrinėtų arba mirusių kalbų gramatikas.

Statistiniais metodais veikiantis lietuvių kalbos sintaksinis analizatorius prieinamas internete kaip *UDPipe* lietuvių kalbos modulis. Jis gerai išanalizuoja tuos lietuvių kalbos sakinius, kuriuose žodžių tvarka yra angliška. Netikslumų analizės metu daugiausia pasitaiko tada, kai sakinys turi specifinių lietuvių kalbos bruožų, pvz., anglų kalbai nebūdingą žodžių išsidėstymą.

Lietuvių kalbai kol kas sukurta tik nedidelė apžvalginė gramatika ir bandomasis skaitmeninės gramatikos pavyzdys.

LIGIS sukūrimo tikslas – parengti lietuvių kalbos gramatikos dokumentaciją, t. y. itin didelio tikslumo ir patikimumo gramatinius duomenis. Siekiama sukurti universalią sistemą, kad ja galėtų naudotis tiek žmonės, tiek kompiuteriai, t. y. kad ji

populiariai pateiktų informaciją plačiajai visuomenei ir būtų tinkama apdoroti lietuvių kalbą kompiuteriu.

Sistema LIGIS plačiajai visuomenei pateikia trijų tipų duomenis apie žodį: morfologinius, morfeminius ir darybinius. Pirmą kartą pateikiama informacija apie morfemos tipą ir jos charakteristikas. Taip pat galima pamatyti kiekvieno žodžio visas kaitybines formas.

Kuriant LIGIS, paaiškėjo naujų faktų apie lietuvių kalbą ir jos gramatiką, į kuriuos anksčiau nebuvo atkreiptas dėmesys, kadangi žmogui, kaip kalbos vartotojui, tie klausimai neaktualūs. Kompiuterizuojant lietuvių kalbos gramatiką pastebėta, kad kai kurie lietuvių kalbos žodžiai negali įgyti visų teorinės paradigmos formų ir šį reiškinį nulemia semantika: taip nutinka dėl leksinės reikšmės nederėjimo su morfologine reikšme arba dėl jų loginio prieštaringumo. Grupinį veiksmą reiškiančių žodžių vienaskaita yra negalima. Įvykio veikslo veiksmažodžių vediniai su priešdėliu *tebe-* negali turėti būtojo kartinio laiko. Neveikiamosios rūšies dalyviai, padaryti iš intranzityvinių veiksmažodžių, gali turėti tik bevardės giminės formą.

Taigi, apdorojant kalbą kompiuteriu, galima pažvelgti į gramatiką iš kitos pusės ir pastebėti žmogui nekrintančią į akis informaciją.

TERMINŲ ŽODYNĖLIS

Terminas	Apibrėžimas
Agliutinacinės kalbos	Tai kalbos, kuriose žodžių formos sudaromos jungiant prie šaknies afiksus. Ir formų kaitybos, ir formų darybos afiksai turi tik vieną apibrėžtą gramatinę reikšmę. Balsių ar priebalsių kaitos šaknyje šiose kalbose nėra.
Algoritmas	Detali veiksmų seka.
ALKSNIS	Anotuotas lietuvių kalbos sintaksinis tekstynas.
Analitinės formos	Veiksmažodžio formos, kai laikams sudaryti naudojami du ar daugiau žodžių. Lietuvių kalboje tai, pvz., *esu matęs*, anglų kalboje *has been used* ir pan.
Analitinės kalbos	Kalbos, kuriose žodžių tarpusavio santykiai daugiausia reiškiami tarnybiniais žodžiais, žodžių tvarka ir kitomis už žodžio ribų iškeltomis gramatinės reikšmės raiškos priemonėmis.
Anotuotas tekstynas	Tekstynas, kuriame sužymėti žodžių morfologiniai ir sintaksiniai duomenys; pastraipų, sakinių ribos; santrumpos, tikriniai daiktavardžiai; garso įrašuose – pauzės, kalbėjimas vienu metu ir kt.
Antecedentas	Žodis, kuris šalutiniame ar toliau po jo einančiame sakinyje pakeistas įvardžiu.
Apibrėžiklis (angl. *determiner*)	Tai artikelis, parodomasis ar asmeninis įvardis ir pan.
Aplinkybiniai pažyminiai	Žodžiai, kuriems galima iškelti ir pažyminio, ir aplinkybės klausimą, pvz.: atostogos *vasarą* (*atostogos kada?* ir *kokios atostogos?*), ežeras *miške* (*ežeras kur?* ir *koks ežeras?*), kelias *atgal* (*kelias kur?* ir *koks kelias?*).
Apžvalginė gramatika	Gramatika, kurios apimtis daug mažesnė už akademinės gramatikos. Joje pateikiamos pagrindinės žinios apie kalbą ir parodomas tos kalbos gramatikos išskirtinumas.
Auksinis standartas	Labai didelio tikslumo ranka žmogaus anotuoti tekstai.
Automatinis mokymasis	Šiuo metu literatūroje dar vartojamas terminas *mašininis mokymasis*.

Terminas	Apibrėžimas
Automatinis vertimas	Šiuo metu literatūroje dar vartojamas terminas *mašininis vertimas*.
Ciklas (grafų teorijoje)	Grafas turi ciklą (ar kelis ciklus), jeigu, išėjus iš kurios nors viršūnės, į ją galima sugrįžti jau kitu keliu.
Dichotominis	Šiame modelyje sakinio branduolį (jo centrą) sudaro veiksnys ir tarinys.
Eliptiniai sakiniai	Sakiniai, kuriuose nėra išreikšto tarinio, bet yra antrininkes sakinio dalis atitinkančių žodžių, pvz., *Lapė į mišką, šunys – paskui*.
Galutiniai simboliai	Formaliųjų gramatikų aprašo dalis: tautų kalboms – tai morfologinės kategorijos ar tiesiog sakinio žodžiai.
Grafas	Grafas – tai grupė objektų, sujungtų linijomis, kurios vadinamos lankais. Patys objektai vadinami viršūnėmis.
Gramatinė struktūra	Skaitmeninėms gramatikoms kurti skirta programinė įranga.
Hunspell (angl. *Hungarian spell checker*)	Vengrų kalbos rašybos klaidų tikrintuvas.
Išplečiantis žodis	Žodis, priklausomas nuo kito žodžio sakinyje, pvz., veiksnio pažyminys yra išplečiantis, nes jis išplečia veiksnį.
Išplečiamas žodis	Žodis, nuo kurio priklauso, prie kurio šliejami ar su kuriuo derinami kiti jį papildantys žodžiai, pvz., veiksnys yra išplečiamas pažyminiu.
Izoliacinės kalbos	Tai kalbos, kurios gali turėti tik labai ribotą morfemų skaičių žodyje. Maksimalios izoliacinės struktūros kalbose žodžiai sudatyti tik iš vienos morfemos – šaknies, nesudaromi nei sudurtiniai žodžiai, nei vediniai su afiksais. Kaitybos nėra, gramatinius ryšius lemia žodžių tvarka sakinyje.
JABLONSKIS	Tekstynams anotuoti skirtos specifinės žymos, kuriose naudojami lietuviškų morfologinių kategorijų sutrumpinimai.
Kolokacija	Iš dalies ar visiškai pastovūs posakiai, kurie tekstyne pasitaiko dažniau, nei tikimasi, pvz., *stipri arbata*, bet *galingas kompiuteris*.
Lapas (grafų teorijoje)	Tai tokia grafo viršūnė, į kurią linija ateina, bet iš jos toliau neišeina nė viena linija.

Terminas	Apibrėžimas
Lema	Žodžio pradinė forma: veiksmažodžiui – bendratis, vardažodžiams – vardininko linksnis.
Lygiagretusis tekstynas	Tai tekstynas, apimantis dviejų ar daugiau kalbų išverstus tekstus ir sulygiuotas pagal pastraipas, sakinius, žodžius.
Medis (grafų teorijoje)	Tai susietas grafas be ciklų.
MULTEXT-East	Žymų rinkinys, naudojamas tekstynams anotuoti, kai anotavimo informacija pateikiama kaip raidžių seka, kurioje kiekviena morfologinė kategorija aprašoma viena raide ir jų išsidėstymas labai griežtas: kiekviena kategorija turi savo poziciją.
Negalutiniai simboliai	Simboliai, nurodantys formalios kalbos sąvokas: tautų kalbose tai – sintaksinės kategorijos, pvz., *daiktavardinė frazė*, *sakinys* ir kt.
Neprojektyvus sakinys	Tai toks sakinys, kurio sintaksinę struktūrą vaizduojant frazių gramatikos medžiu, atsiranda susikertančių linijų.
Pakeitimo taisyklė	Formalių gramatikų taisyklė, nurodanti, kad tam tikras simbolis medyje gali būti išskleistas kitų simbolių seka.
Ontologija	Pagrindinis filosofijos skyrius – būties mokslas, nagrinėjantis būties teoriją, tikrovės pobūdį ir struktūrą.
PAULA	Tekstynų anotavimo formatas, kuriame kiekvieną morfologinę kategoriją žymi simbolių grupė, apribota iš abiejų pusių taškais.
Polisintetinės kalbos	Dar vadinamos *inkorporacinės kalbos*, kuriose sakiniai sudaromi kaip sudėtiniai žodžiai, t. y. šaknys jungiamos į bendrą visumą, į žodį-sakinį. Pvz., čiukčių kalboje žodžio-sakinio pradžią sudaro veiksnio šaknis, pabaigą – tarinio. Tarp jų įterpiamos papildinių su pažyminiais ir aplinkybių šaknys.
Pradinis simbolis	Frazių gramatikoje tai negalutinis simbolis S, nuo kurio pradedamas eilučių generavimas arba kuris turi būti gaunamas analizės metu.
Projektyvus sakinys	Sakinys, kurio sintaksinę struktūrą frazių gramatikoje galima pavaizduoti be susikertančių linijų.
Sentimentų analizatorius	Programinė įranga, klasifikuojanti tekstus (dažniausiai interneto puslapių komentarus) pagal teigiamą, neigiamą ir neutralų emocinį atspalvį.

Terminas	Apibrėžimas
Sintetinės kalbos	Kalbos, kuriose žodžių tarpusavio santykiai daugiausia reiškiami gramatinėmis morfemomis – galūnėmis, priesagomis, priešdėliais, pvz., lietuvių kalba.
Susietas grafas	Grafas vadinamas susietu, jei tarp bet kurių dviejų jo viršūnių egzistuoja kelias.
Šaknis (grafų teorijoje)	Šaknis yra tokia grafo viršūnė, į kurią neateina jokia linija, iš jos linijos gali tik išeiti.
Tarnybiniai žodžiai	Nekaitomi nesavarankiški žodžiai, neįvardijantys daiktų, veiksmų ar požymių.
Tautų kalbos	Šiuo metu literatūroje dar vartojamas terminas *natūralios kalbos*.
Tekstynas	Didelis kiekis tekstų, sukauptų elektronine forma, kurie naudojami mokslinių tyrimų tikslams ir analizuojami specialiomis programinėmis priemonėmis.
Valentingumas	„...žodžių ir jų formų gebėjimas sakinyje jungtis su kitomis žodžių formomis. Lietuvių kalbos sakinio sandarai svarbiausias veiksmažodžių valentingumas...“ (Ambrazas 2022). Kai kurie veiksmažodžiai reikalauja vieno linksnio, pvz., *vaikas miega*, kiti – dviejų, pvz., *sesuo rašo laišką*, ar trijų ir daugiau, pvz., *mergaitė padovanojo draugei knygą*.
Verbocentrinis	Šiame modelyje sakinys turi tik vieną centrą – tarinį, skirtingai nuo tradicinės gramatikos, kur sakinio branduolį sudaro veiksnys ir tarinys.

SANTRUMPOS

Santrumpa	Išskleidimas
BLEU	Angl. *Bilingual evaluation understudy* – automatinio vertimo kokybės įverčiai, kai skaičiuojama, kokią dalį sakinių kompiuteris išvertė teisingai, palyginti su žmogaus vertimu.
BPR	Angl. *Bayesian Personalized Ranking* – taikomasis išrikiavimas remiantis tikėtinumu, apskaičiuotu pagal Tomo Beizo (Thomas Bayes) teoremą. Tai alternatyvus (lyginant su tradicine, dažnumais besiremiančia statistika) tikėtinumo apskaičiavimo būdas.
CoNLL-U	Angl. *Computational Natural Language Learning – Universal Dependencies,* tai – duomenų formatas, naudojamas automatinio mokymo tikslams, kai sakinio struktūra pavaizduojama universaliųjų priklausomybių metodu.
CPA	Angl. *Corpus Pattern Analysis* – tekstyno modelių analizė.
det	Angl. *determiner* – apibrėžiklis.
DGS	Vok. ***D****eutsche* ***G****ebärden****s****prache* – vokiečių gestų kalba.
DLKT	Dabartinės lietuvių kalbos tekstynas.
DUG	Angl. *Dependency Unification Grammar* – formalus gramatikos aprašas.
ELRC	Angl. *European Language Resource Coordination* – Europos kalbų išteklių koordinavimas.
LA	*Los Angeles*.
LGR	*Leipzig Glossing Rules* – žymų standartas, naudojamas tekstynams anotuoti.
LIGIS	**Li**etuvių kalbos **g**ramatikos **i**nformacinė **s**istema.
LKG	Lietuvių kalbos gramatika.
n	Angl. *noun* – daiktavardis.
NP	Angl. *Noun Phrase* – daiktavardinė frazė.
PML	*Prague Markup Language* – Čekijoje sukurta metodika, skirta tekstynams anotuoti.
POS	Angl. *Part Of Speech* – kalbos dalis.
PP	Angl. *Prepositional Phrase* – prielinksninė frazė (prielinksninė konstrukcija).

Santrumpa	Išskleidimas
prep	Angl. *preposition* – prielinksnis.
pron	Angl. *pronoun* – įvardis.
RGL	Angl. *Resource Grammar Library – gramatikos išteklių biblioteka.*
S	Angl. *Sentence* – sakinys.
SGR	*Salos Glossing Rules* – baltų kalboms sudarytas žymų rinkinys, skirtas tekstynams anotuoti.
SOV	Žodžių tvarka sakinyje: *Subject–Object–Verb* (veiksnys–papildinys–tarinys).
SOVQ	Klausiamojo sakinio žodžių tvarka: *Subject–Object–Verb–Question*.
SVO	Žodžių tvarka sakinyje: *Subject–Verb–Object* (veiksnys–tarinys–papildinys).
SVOQ	Klausiamojo sakinio žodžių tvarka: *Subject–Verb–Object–Question*.
Tab-WPL	Angl. *Tab delimited Word Per Line* – formatas, kuriame kiekvienas sakinio žodis rašomas į atskirą eilutę ir žodžio duomenų tipai eilutėje atskiriami tabuliacijomis
UDPipe	Angl. *Universal Dependencies Pipeline* – Universaliųjų priklausomybių konvejeris (programinė įranga, atliekanti morfologinę analizę bei sintaksinę analizę universaliųjų priklausomybių metodu).
v	Angl. *verb* – veiksmažodis.
VOS	Žodžių tvarka sakinyje: *Verb–Object–Subject* (tarinys–papildinys–veiksnys).
VP	Angl. *Verb Phrase* – veiksmažodinė frazė.
XML	Angl *Extensible Markup Language* – bendros paskirties duomenų struktūrų aprašymo formatas.

INTERNETO NUORODOS

1 interneto nuoroda: Dėl Lietuvių kalbos plėtros skaitmeninėje terpėje ir kalbos technologijų pažangos 2021–2027 metų gairių patvirtinimo. *Dokumentų paieška.* https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/911407f20ee911ebbedbd456d2fb030d?positionInSearchResults=0&searchModelUUID=a5713c95-b943-4db5-81c0-84b7c1d74208&fbclid=IwAR0IS_88QUeoq_Cv_7XaLj0mIWoHqu7JOarz1AaCn4MfqIYogOQafELXOmk [žiūrėta 2022-11-22].

2 interneto nuoroda: Cutting up words: Introduction to morphology, morphemes, stems, prefixes and suffixes. *YouTube.* https://www.youtube.com/watch?v=8ypQq5MvT24 [žiūrėta 2022-11-22].

3 interneto nuoroda: Švedų kalba. *WikipediA*. https://lt.wikipedia.org/wiki/%C5%A0ved%C5%B3_kalba [žiūrėta 2022-11-22].

4 interneto nuoroda: Tamsioji dirbtinio intelekto paslaptis: niekas iš tiesų nesupranta, kaip jis veikia. *Technologijos.lt*, 2017-05-02. http://www.technologijos.lt/n/technologijos/it/S-61326/straipsnis/Tamsioji-dirbtinio-intelekto-paslaptis-niekas-is-tiesu-nesupranta-joveikimo [žiūrėta 2022-11-22].

5 interneto nuoroda: Konvoliucinis neuroninis tinklas. *WikipediA*. https://lt.wikipedia.org/wiki/Konvoliucinis_neuroninis_tinklas [žiūrėta 2022-11-22].

6 interneto nuoroda: *Parts-of-speech.Info.* https://parts-of-speech.info/ [žiūrėta 2022-11-22].

7 interneto nuoroda: *Wortarten.Info.* https://wortarten.info/ [žiūrėta 2022-11-22].

8 interneto nuoroda: Beribis lietuvių kalbos pasaulis skaitmeninių išteklių sistemoje „E.kalba“. *Pasaulio lietuvis.* https://pasauliolietuvis.lt/beribis-lietuviu-kalbos-pasaulis-skaitmeniniu-istekliu-sistemoje-e-kalba/ [žiūrėta 2022-11-22].

9 interneto nuoroda: Lietuvių kalbos išteklių informacinės sistemos „E. kalba“ paslauga „Nuomonių analizė“. *E.KALBA.* https://ekalba.lt/public#/sentimentAnalysis/about [žiūrėta 2020-04-17].

10 interneto nuoroda: VDU mokslininkai vysto dirbtinio intelekto technologijų sprendimus lietuvių kalbai: kodai bus perduoti visuomenei. *Lietuvos Aidas*, 2018-09-27. http://www.aidas.lt/lt/mokslas-ir-it/article/22560-09-27-vdu-mokslininkai-vysto-dirbtinio-intelekto-technologiju-sprendimus-lietuviu-kalbai-kodai-bus-perduoti-visuomenei [žiūrėta 2022-11-22].

11 interneto nuoroda: Kalbos technologijos – būtina sąlyga kalbai gyvuoti. *Alkas.lt*, 2017-11-03. http://alkas.lt/2017/11/03/kalbos-technologijos-butina-salyga-kalbai-gyvuoti/ [žiūrėta 2022-11-22].

12 interneto nuoroda: Machine translation for public administrations – eTranslation. *Europos Komisija*. https://ec.europa.eu/info/resources-partners/machine-translation-public-administrations-etranslation_en [žiūrėta 2022-11-22].

13 interneto nuoroda: Evaluating models. *Google Cloud*. https://cloud.google.com/translate/automl/docs/evaluate [žiūrėta 2022-11-22].

14 interneto nuoroda: Textcorpus. *WikipediA*. https://en.wikipedia.org/wiki/Text_corpus. [žiūrėta 2022-11-22].

15 interneto nuoroda: Apie Dabartinės lietuvių kalbos tekstyną. *KLC*. http://tekstynas.vdu.lt/tekstynas/menu?page=about [žiūrėta 2022-11-22].

16 interneto nuoroda: Textkorpus. *WikipediA*. https://de.wikipedia.org/wiki/Textkorpus [žiūrėta 2022-11-22].

17 interneto nuoroda: Corpus of Contemporary American English. *English-Corpora.org*. https://www.english-corpora.org/coca/ [žiūrėta 2022-11-22].

18 interneto nuoroda: *English-Corpora.org*. https://www.english-corpora.org/ [žiūrėta 2022-11-22].

19 interneto nuoroda: Diktoriaus skaitomo rišlaus teksto garsynas. *VDU CRIS*. https://www.vdu.lt/cris/handle/20.500.12259/41313?mode=full [žiūrėta 2022-11-22].

20 interneto nuoroda: Užimtumas, socialiniai reikalai ir įtrauktis: Mūsų kalbų politika. *Europos Komisija*. https://ec.europa.eu/social/main.jsp?catId=521&langId=lt [žiūrėta 2022-11-22].

21 interneto nuoroda: Rozetos akmuo. *WikipediA*. https://lt.wikipedia.org/wiki/Rozetos_akmuo [žiūrėta 2022-11-22].

22 interneto nuoroda: Rozetės akmuo. *Istorija Tau*. https://istorijatau.lt/rubrikos/zodynas/rozetes-akmuo [žiūrėta 2022-11-22].

23 interneto nuoroda: POStags: What is a POStag? *Sketch Engine*. https://www.sketchengine.eu/pos-tags/ [žiūrėta 2022-11-22].

24 interneto nuoroda: Parts-of-speech tags used. Brown Corpus. *WikipediA*. https://en.wikipedia.org/wiki/Brown_Corpus#Part-of-speech_tags_used [žiūrėta 2022-11-22].

25 interneto nuoroda: MATAS – morfologiškai anotuotas tekstynas. *KLC*. https://klc.vdu.lt/matas-morfologiskai-anotuotas-tekstynas/ [žiūrėta 2022-11-22].

26 interneto nuoroda: Morphology: General Principles. *Univwesal Dependencies*. https://universaldependencies.org/u/overview/morphology.html [žiūrėta 2022-11-22].

27 interneto nuoroda: Paieška tekstyne. DLKT. *Paieška*. http://corpus.vdu.lt [žiūrėta 2022-11-22].

28 interneto nuoroda: Paieška tekstyne – dangaus. DLKT. *Paieška*. http://corpus.vdu.lt/lt/?word=dangaus. [žiūrėta 2022-11-22]

29 interneto nuoroda: Brown Corpus. *WikipediA*. https://en.wikipedia.org/wiki/Brown_Corpus [žiūrėta 2022-11-22].

30 interneto nuoroda: The SUSANNE Corpus: Documentation – 4.1. Field Structure. *Geoffrey Sampson*. https://www.grsampson.net/SueDoc.html [žiūrėta 2022-11-22].

31 interneto nuoroda: UCREL CLAWS7 Tagset. *UCREL Lancaster UK*. http://ucrel.lancs.ac.uk/claws7tags.html [žiūrėta 2022-11-22].

32 interneto nuoroda: Reading Negra corpus files with a Negra Corpus Reader. *Git Hub*. https://github.com/nltk/nltk/issues/137 [žiūrėta 2022-11-22] .

33 interneto nuoroda: Verse (67:1) – Quranic Syntax. *The Quranic Arabic Corpus*. http://corpus.quran.com/treebank.jsp?chapter=67 [žiūrėta 2022-11-22].

34 interneto nuoroda: *The Ancient Greek and Latin Dependency Treebank*. http://nlp.perseus.tufts.edu/syntax/treebank/ [žiūrėta 2012-12-20].

35 interneto nuoroda: Prague Dependency Treebank 2.0: Layers of Annotation. *UFAL*. http://ufal.mff.cuni.cz/pdt2.0/doc/pdt-guide/en/html/ch02.html [žiūrėta 2022-11-22].

36 interneto nuoroda: ALKSNIS – sintaksiškai anotuotas tekstynas. *KLC*. https://klc.vdu.lt/alksnis-sintaksiskai-anotuotas-tekstynas/ [žiūrėta 2022-11-22].

37 interneto nuoroda: Rengiamas lietuvių kalbos sintaksiškai anotuotas tekstynas ALKSNIS. *CLARIN-LT*. http://clarin-lt.lt/?p=205 [žiūrėta 2022-11-22].

38 interneto nuoroda: Patricia Trie. *WikipediA*. https://de.wikipedia.org/wiki/Patricia-Trie [žiūrėta 2022-11-22].

39 interneto nuoroda: Radix tree. *WikipediA*. http://en.wikipedia.org/wiki/Radix_tree [žiūrėta 2022-11-22].

40 interneto nuoroda: Morfologinis anotatorius. *KLC*. https://klc.vdu.lt/anotatorius/ [žiūrėta 2022-11-22].

41 interneto nuoroda: Paieška tekstyne – nebeatsinešdavau. DLKT. *Paieška*. http://corpus.vdu.lt/lt/?word=nebeatsine%C5%A1davau [žiūrėta 2022-11-22].

42 interneto nuoroda: Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema. *web.archive.org*. https://web.archive.org/web/20200221090527/http://www.semantika.lt:80/SyntaticAndSemanticAnalysis/Analysis. [žiūrėta 2022-11-22].

43 interneto nuoroda: Lietuviško teksto analizė ir taisymas. *Lietuvių kalbos sintaksinės ir semantinės analizės informacinė sistema*. http://www.semantika.lt/SyntaticAndSemanticAnalysis/Analysis [žiūrėta 2016-10-26].

44 interneto nuoroda: Lietuviško teksto analizė ir taisymas. *SEMANTIKA.LT*. https://www.semantika.lt/Analysis/TextAnalysis [žiūrėta 2022-11-22].

45 interneto nuoroda: LINDAT/CLARIN Servises: UD Pipe. *LINDATCLARIAH-CZ*. https://lindat.mff.cuni.cz/services/udpipe/run.php [žiūrėta 2022-11-22].

46 interneto nuoroda: Wordstructure. *Slide Share*. https://www.slideshare.net/riortamm/word-structure-42667712 [žiūrėta 2022-11-22].

47 interneto nuoroda: Краткая русская грамматика: Морфемная структура слова. Виды морфем. – Словоформы разнообразной морфемной структуры. *Langust*. https://www.langust.ru/rus_gram/rus_gr03.shtml [žiūrėta 2022-11-22].

48 interneto nuoroda: Морфемное членение. *old.kpfu.ru*. http://old.kpfu.ru/infres/slovar1/begall.htm [žiūrėta 2021-12-02].

49 interneto nuoroda: Free English Morphological Parsing Service: An English Morphological Parser. *Nlpdotnet*. http://nlpdotnet.com/services/Morphparser.aspx [žiūrėta 2016-01-22].

50 interneto nuoroda: Lietuvių kalbos morfemikos duomenų bazė. *CLARIN-LT*. https://klc.vdu.lt/morfema/ [žiūrėta 2022-11-22].

51 interneto nuoroda: Garden-path sentence. *WikipediA*. https://en.wikipedia.org/wiki/Garden-path_sentence [žiūrėta 2022-11-22].

52 interneto nuoroda: Universal Dependencies V1 Documentation: Introduction. *Universal Dependencies*. http://universaldependencies.org/docsv1/introduction.html [žiūrėta 2022-11-22].

53 interneto nuoroda: Sintaksiškai anotuoto tekstyno ALKSNIS naudojimas. *YouTube*. https://www.youtube.com/watch?v=PlE0PWurb4Y&t=29s 30-ta sekundė [žiūrėta 2022-11-22].

54 interneto nuoroda: Statistical parsing. *WikipediA*.
https://en.wikipedia.org/wiki/Statistical_parsing
[žiūrėta 2022-11-22].

55 interneto nuoroda: Can-can. *WikipediA*.
https://en.wikipedia.org/wiki/Can-can [žiūrėta 2022-11-22].

56 interneto nuoroda: Parsing. *WikipediA*.
https://en.wikipedia.org/wiki/Parsing [žiūrėta 2022-11-22].

57 interneto nuoroda: SpaCy. *WikipediA*.
https://en.wikipedia.org/wiki/SpaCy [žiūrėta 2022-11-22].

58 interneto nuoroda: Models-Lithuanian. *SpaCy*.
https://spacy.io/models/lt [žiūrėta 2022-11-22].

59 interneto nuoroda: *karabatos.gr*.
https://www.karabatos.gr/ [žiūrėta 2022-11-22].

60 interneto nuoroda: Meine Grammatik – digital. *karabatos.gr*.
http://www.karabatos.gr/de/meine-grammatik-digital-cd-rom-f%-C3%BCr-interaktive-whiteboards [žiūrėta 2022-11-22].

61 interneto nuoroda: *Deutsch – Digital*.
http://deutsch-digital.nl/index_grammatica.htm
[žiūrėta 2014-05-30].

62 interneto nuoroda: *Digital grammatik*.
http://digitalgrammatik.blogspot.com/ [žiūrėta 2022-11-22].

63 interneto nuoroda: Word sketch. *WikipediA*.
https://en.wikipedia.org/wiki/Word_sketch [žiūrėta 2022-11-22].

64 interneto nuoroda: What can Sketch Engine do with a Word? *Sketch Engine*.
https://www.sketchengine.eu/what-can-sketch-engine-do/
[žiūrėta 2022-11-22].

65 interneto nuoroda: Word Sketch – collocations and word combinations. *Sketch Engine*.
https://www.sketchengine.eu/guide/word-sketch-collocations-and-word-combinations/ [žiūrėta 2022-11-22].

66 interneto nuoroda: Writing a Sketch Grammar: Grammatical Relation Definitions. *Sketch Engine*.
https://www.sketchengine.eu/documentation/writing-sketch-grammar/ [žiūrėta 2022-11-22].

67 interneto nuoroda: Terminų žodynas: Indukcija. *Žodynas.lt*.
https://www.zodynas.lt/terminu-zodynas/I/indukcija
[žiūrėta 2022-11-22].

68 interneto nuoroda: Projektas „Užsienio baltistikos centrų ir Lietuvos mokslo ir studijų institucijų bendradarbiavimo skatinimas“. *Baltnexus*.
https://baltnexus.lt/lt/baltistikos-projektas skirtukas *3.2.2. Interaktyvios mokymo priemonės* [žiūrėta 2022-11-22].

69 interneto nuoroda: *Kalbu*.
https://kalbu.vdu.lt/
[žiūrėta 2022-11-22].

70 interneto nuoroda: Technology. *digital Grammars*.
https://www.digitalgrammars.com/technology
[žiūrėta 2022-11-22].

71 interneto nuoroda: Grammatical Framework. *WikipediA*.
https://en.wikipedia.org/wiki/Grammatical_Framework
[žiūrėta 2022-11-22].

72 interneto nuoroda: GF Resource Grammar Library: Synopsis. *Grammatical Framework*.
http://www.grammaticalframework.org/lib/doc/synopsis/
[žiūrėta 2022-11-22].

73 interneto nuoroda: Izoliacinė kalba. *WikipediA*.
https://lt.wikipedia.org/wiki/Izoliacin%C4%97_kalba
[žiūrėta 2022-11-22].

74 interneto nuoroda: Minigrammar LIT. *GF online editor for simple multilingual grammars*.
http://cloud.grammaticalframework.org/gfse/
[žiūrėta 2022-11-22].

75 interneto nuoroda: *Google* vertėjas – pelę mato katė?. Translate. *Google*.
https://translate.google.com/?sl=lt&tl=en&text=pel%C4%99%20mato%20kat%C4%97%3F&op=translate [žiūrėta 2022-11-22].

76 interneto nuoroda: *LIGIS – Lietuvių kalbos gramatikos informacinė sistema*.
http://ligis.lki.lt/ [žiūrėta 2022-11-22].

77 interneto nuoroda: susitikimas – gramatinės formos. *MORFOLOGIJA.LT*.
https://morfologija.lietuviuzodynas.lt/zodzio-formos/susitikimas
[žiūrėta 2022-11-22].

78 interneto nuoroda: Analizė – nubėgti. *LIGIS*.
http://ligis.lki.lt/?wordInput=nub%C4%97gti
[žiūrėta 2022-11-22].

79 interneto nuoroda: Морфологический разбор слова онлайн. *Goldlit*.
https://goldlit.ru/component/slog?words=%D0%BF%D0%BE%D0%B4%D0%B3%D0%BE%D1%82%D0%BE%D0%B2%D0%BB%D0%B5%D0%BD%D0%B0 [žiūrėta 2022-11-22].

80 interneto nuoroda: Analizė – bėgtakis – Kitos formos. *LIGIS*.
http://ligis.lki.lt/?wordInput=b%C4%97gtakis
[žiūrėta 2022-11-22].

81 interneto nuoroda: Analizė – bėgį – Kitos formos. *LIGIS*.
http://ligis.lki.lt/?wordInput=b%C4%97g%C4%AF
[žiūrėta 2022-11-22].

82 interneto nuoroda: Pioneering the computational linguistics and the largest published work of alltime. *web.archive.org*.
https://web.archive.org/web/20120327122219/http://www.ibm.com/ibm100/it/en/stories/linguistica_computazionale.html
[žiūrėta 2022-11-22].

83 interneto nuoroda: POS tags. *Sketch Engine*.
https://www.sketchengine.eu/blog/pos-tags/ [žiūrėta 2022-11-22].

84 interneto nuoroda: The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses. *ewa.mpg.de*. https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf [žiūrėta 2022-11-22].

85 interneto nuoroda: Maratono treniruotės pradedantiesiems. *dovydas.sankauskas.lt* http://www.dovydas.sankauskas.lt/wiki/maratono_treniruot%C4%97s_pradedantiesiems_10_savaitė [žiūrėta 2018-03-03].

86 interneto nuoroda: Įvertinkite pokyčius: kardinaliai stilių pakeitusi Rosita Čivilytė pristatė dainą „Man gana“. *lrytas.lt*. https://www.lrytas.lt/zmones/muzika/2020/06/29/news/ivertinkite-pokycius-kardinaliai-stiliu-pakeitusi-rosita-civilyte-pristate-daina-man-gana--15441488/ [žiūrėta 2022-11-22].

87 interneto nuoroda: Tamil language. *WikipediA*. https://en.wikipedia.org/wiki/Tamil_language [žiūrėta 2022-11-22].

88 interneto nuoroda: Part of Speech. *WikipediA*. https://en.wikipedia.org/wiki/Part_of_speech [žiūrėta 2022-11-22].

89 interneto nuoroda: Zahlwort. *WikipediA*. https://de.wikipedia.org/wiki/Zahlwort [žiūrėta 2022-11-22].

90 interneto nuoroda: Japanese equivalents of adjectives. *WikipediA*. https://en.wikipedia.org/wiki/Japanese_equivalents_of_adjectives [žiūrėta 2022-11-22].

91 interneto nuoroda: Panini (Grammatiker). *WikipediA*. https://de.wikipedia.org/wiki/Panini_(Grammatiker) [žiūrėta 2022-11-22].

92 interneto nuoroda: Yaska. *WikipediA*. https://en.wikipedia.org/wiki/Y%C4%81ska [žiūrėta 2022-11-22].

93 interneto nuoroda: Lingvistika: Kalbotyra Romos laikais. *Mokslai.lt*. https://mokslai.lt/referatai/lietuviu-kalba/lingvistika.html [žiūrėta 2022-11-22].

94 interneto nuoroda: Morfologinio nagrinėjimo tvarka. *Šaltiniai*. http://www.šaltiniai.info/index/details/599%20 [žiūrėta 2022-11-22].

95 interneto nuoroda: Kalbos dalis. *WikipediA*. https://lt.wikipedia.org/wiki/Kalbos_dalis [žiūrėta 2022-11-22].

96 interneto nuoroda: Das deutsche Verb: drei Grundformen der Verben. *LingQ*. https://www.lingq.com/lesson/25-drei-grundformen-der-verben-468764/ [žiūrėta 2022-11-22].

97 interneto nuoroda: Verbs: the three basic forms. *Cambridge Dictionary*. http://dictionary.cambridge.org/grammar/british-grammar/about-verbs/verbs-basic-forms [žiūrėta 2022-11-22].

98 interneto nuoroda: Žemėj Lietuvos. *Musixmatch*. https://www.musixmatch.com/es/letras/Thundertale/%C5%BDem%C4%97j-Lietuvos [žiūrėta 2020-12-21].

99 interneto nuoroda: Teisėjas Darius Kantaravičius gelbėja narkodilerius. *Laisvas Laikraštis*. https://laisvaslaikrastis.lt/kaip-teisejas-darius-kantaravicius-pridengineja-narkodilerius/ [žiūrėta 2019-10-10].

100 interneto nuoroda: Žodis – spiesti. *LKŽ* http://www.lkz.lt/?zodis=spiesti&id=24168780000 [žiūrėta 2022-11-22].

LITERATŪRA

A

Adamski Marcin, Zimniewicz Michal 2011: Automatic Syntactic Analysis for Polish Language. – *Proceedings of the 5th Language & Technology Conference*, Poznan, 536–540. Prieiga internete: https://www.researchgate.net/publication/311674985_Automatic_Sentiment_Analysis_in_Polish_Language [žiūrėta 2022-11-22].

Adedimeji Mahfouz Adebola 2005: Word Structure in English. – *Basic Communication Skills for Students of Science and Humanities*, Ilorin, 1–21.

Agel Vilmos 2000: *Valenztheorie*, Tübingen: Narr.

Aleksa Melita 2006: Automatic Morphological Analysis of the Croatian Language. – *CESCLI – Proceedings of the First Central European Student Conference in Linguistics*, 1–18. Prieiga internete: http://www.nytud.hu/cescl/proceedings/Melita_Aleksa_CESCL.pdf [žiūrėta 2022-11-22].

Allen James 1987: *Natural Language Understanding*, Amsterdam: The Benjamin / Cummings Publishing Company.

Al-Onaizan Yaser, Curun Jan, Jahr Michael, Knight Kevin, Lafferty John, Melamed Dan, Och Franz-Joseph, Purdy David, Smith Noah, Yarowsky David 1999: Statistical Machine Translation. – *Final Report JHU Workshop*, 1–12.

Ambrazas Vytautas 2022: Valentingumas. – *Visuotinė lietuvių enciklopedija*, Vilnius: Mokslo ir enciklopedijų leidybos centras. Prieiga internete: https://www.vle.lt/straipsnis/valentingumas-1/ [žiūrėta 2022-11-22].

Ambrazas Vytautas (red.) 2006: *Lithuanian Grammar*, Vilnius: „Baltos lankos“.

Ambrazas Vytautas (red.) 1999: *Lietuvių kalbos enciklopedija*, Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Ambrazas Vytautas (red.) 1997: *Dabartinės lietuvių kalbos gramatika*, Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Ambrazas Vytautas 1979: *Lietuvių kalbos dalyvių istorinė sintaksė*, Vilnius: „Mokslas“.

Ananiadou Sophia 1987: A brief survey of some current operational systems. – S. Michaelson, Y. Wilks (series Ed.), *Machine Translation Today: The State of The Art. – Information Technology Series*, Edinburgh: Edinburgh University Press, 171–191.

Aputis Juozas 1977: *Sugrįžimas vakarėjančiais laukais*: novelės, Vilnius: „Vaga“.

Arkadiev Peter 2010: Notes on the Lithuanian restrictive. – *Baltic Linguistics* 1, 9–49. Prieiga internete: https://inslav.ru/images/stories/people/arkadiev/Arkadiev_Lithuanian_te_BL2010.pdf [žiūrėta 2022-11-22].

Arnold Doug, Balkan Lorna, Meijer Siety, Humphreys Lee, Sadler Louisa 1994: *Machine Translation: An Introductory Guide*, Cambridge: Blackwell Publishers.

B

Baker Mark 2003: *Lexical categories: verbs, nouns and adjectives*, Cambridge: Cambridge University Press.

Balčikonis Juozas, Larinas Borisas, Kruopas Jonas 1957: *Pirmoji lietuvių kalbos gramatika*, Vilnius: Valstybinė politinės ir mokslinės literatūros leidykla.

Bamman David, Crane Gregory 2011: The Ancient Greek and Latin Dependency Treebanks. – *Language Technology for Cultural Heritage*, 79–98. Prieiga internete: http://nlp.perseus.tufts.edu/docs/latech.pdf [žiūrėta 2022-11-22].

Barsky Robert 2017: Universal Grammar. – *Encyclopaedia Britannica*, Encyclopaedia Britannica inc. Prieiga internete: https://www.britannica.com/topic/universal-grammar [žiūrėta 2022-11-22].

Barzdiņš Guntis, Gosko Didzis 2016: RIGA at SemEval-2016 Task 8: Impact of Smatch Extensions and Character-Level Neural Translation on AMR Parsing Accuracy. – *Proceedings of Sem Eval-2016*, San Diego, California, June 16–17, 2016, Association for Computational Linguistics, 1143–1147. Prieiga internete: https://aclweb.org/anthology/S/S16/S16-1176.pdf [žiūrėta 2022-11-22].

Barzdiņš Guntis, Grūzītis Normunds, Nešpore Gunta, Saulīte Baiba, Auziņa Ilze, Levāne-Petrova Kristīne 2008: μ-Ontologies: Integration of Frame Semantics and Ontological Semantics. – *Proceedings if the 13th EURALEX International Congress*, Barcelona, 277–283. Prieiga internete: https://www.researchgate.net/publication/267793694_Multidimensional_Ontologies_Integration_of_Frame_Semantics_and_Ontological_Semantics [žiūrėta 2022-11-22].

Batori Istvan, Lenders Winfried, Putschke Wolfgang 1989: *Computerlinguistik: Ein internationales Handbuch zur computergestützten Sprachforschung und ihrer Anwendungen*, Berlin: Walter de Gruyter.

Berral Jozep LI, Goiri Inigo, Nou Ramon, Julia Ferran, Guitart Jordi, Gavalda Ricard, Torres Jordi 2010: Towards energy-aware scheduling in data centers using machine learning. – *e-Energy '10: Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking*, April 2010, 215–224. Prieiga internete: https://www.researchgate.net/publication/221561415_Towards_energy-aware_scheduling_in_data_centers_using_machine_learning/figures?lo=1 [žiūrėta 2022-11-22].

Bielinskienė Agnė, Boizou Loic, Kovalevskaitė Jolanta, Rimkutė Erika 2016: Lithuanian Dependency Treebank ALKSNIS. – *Human Language Technologies – The Baltic Perspective*, I. Skadiņa and R. Rozis (eds.), 107–114. Prieiga internete: https://etalpykla.lituanistikadb.lt/object/LT-LDB-0001:J.04~2016~1569935723324/J.04~2016~1569935723324.pdf [žiūrėta 2022-11-22].

Bielinskienė Agnė, Boizou Loic, Rimkutė Erika 2017: Lietuvių kalbos morfologiškai ir sintaksiškai anotuoti tekstynai. – *Bendrinė kalba* 90, 1–30. Prieiga internete: http://www.bendrinekalba.lt/Straipsniai/90/Bielinskiene%20ir%20kt_BK_90_straipsnis.pdf [žiūrėta 2022-11-22].

Boguslavsky Igor, Grigirieva Svetlana, Grigoriev Nikolai, Kreidlin Leonid, Frid Nadezhda 2011: *Dependency Treebank for Russian: Concept, Tools, Types of Information*. Prieiga internete: https://www.researchgate.net/publication/221101925_Dependency_Treebank_for_Russian_Concept_Tools_Types_of_Information#fullTextFileContent [žiūrėta 2022-11-22].

Boizou Loic, Kapočiūtė-Dzikienė Jurgita, Rimkutė Erika 2018: Deeper Error Analysis of Lithuanian Morphological Analyzers. – *Human Language Technologies – The Baltic Perspective*, K. Muischnek and K. Müürisep (eds.), 18–25. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/59510/2/ISBN9781614999119.PG_18-25.pdf [žiūrėta 2022-11-22].

Boizou Loic, Kovalevskaitė Jolanta, Rimkutė Erika 2020: Lithuanian Pedagogic Corpus: Correlations Between Linguistic Features and Text Complexity. – *Human Language Technologies – The Baltic Perspective*, A. Utka et al. (eds.), 233–240. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/110434/2/ISBN9781643681177.PG_233-240.pdf [žiūrėta 2022-11-22].

Boizou Loic, Zamblera Francesco 2014: Syntactic Engine for the Lithuanian Language. – *Human Language Technologies – The Baltic Perspective: Proceedings of the 6th International Conference Baltic HLT*, A. Utka et al. (eds.), Amsterdam: IOS Press, 69–74. Prieiga internete: https://ebooks.iospress.nl/volumearticle/38006 [žiūrėta 2022-11-22].

Bosch van den Antal, Daelemans Walter 1999: Memory-Based Morphological Analysis. – *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*, 285–292. Prieiga internete: https://dl.acm.org/doi/pdf/10.3115/1034678.1034726 [žiūrėta 2022-11-22].

Brown Peter, Cocke John, Della Pietra, Stephen Andrew, Della Pietra, Vincent Joseph, Jelineck Frederick, Lafferty John, Mercer Robert, Roossin Paul 1990: A Statistical Approach to Machine Translation. – *Computational Linguistics* 16(2), 79–85.

Buch-Kromann Matthias 2010: The DTAG treebank tool. – *Annotating and querying treebanks and parallel treebanks / Working paper*. Prieiga internete: https://github.com/mbkromann/copenhagen-dependency-treebank/blob/master/docs/2010-wp-dtag.pdf [žiūrėta 2022-11-22].

Bungeroth Jan, Stein Daniel, Dreuw Philippe, Zahedi Morteza, Ney Hermann 2006: *A German Sign Language Corpus of the Domain Weather Report*. Prieiga internete: https://www.cs.brandeis.edu/~marc/misc/proceedings/lrec-2006/pdf/673_pdf.pdf [žiūrėta 2022-11-22].

C

Caswell Isaak 2022: *Google Translate learns 24 new languages*. Prieiga internete: https://www.blog.google/products/translate/24-new-languages/ [žiūrėta 2022-11-22].

Chabris Christofer 1989: *Artificial Intelligence & Turbo C*, Homewood: Dow Jones-Irwin.

Charniak Eugene 1997: Statistical Techniques for Natural Language Parsing. – *AI Magazine* 18(4), 33–43. Prieiga internete: https://ojs.aaai.org//index.php/aimagazine/article/view/1320 [žiūrėta 2022-11-22].

Choi Jinho 2016: Dynamic Feature Induction: The Last Gist to the State-of-the-Art. – *Proceedings of NAACL-HLT 2016*, San Diego, California, June 12–17, Association for Computational Linguistics, 271–281. Prieiga internete: https://www.researchgate.net/publication/305334568_Dynamic_Feature_Induction_The_Last_Gist_to_the_State-of-the-Art [žiūrėta 2022-11-22].

Chomsky Noam 1993: *Lectures on government and binding*, Berlin: Walter de Gruyter & Co.

Chomsky Noam 1956: Three models for the description of language. – *IRE Transactions on Information Theory* 2, 113–124. Prieiga internete: https://ieeexplore.ieee.org/document/1056813 [žiūrėta 2022-11-22].

Chung Sandra 2012: Are Lexical categories universal? The view from Chamorro. – *Theoretical linguistics* 38(1–2), 1–56. Prieiga internete: https://www.researchgate.net/publication/274908779_Are_lexical_categories_universal_The_view_from_Chamorro [žiūrėta 2022-11-22].

Creutz Mathias, Lagus Krista, Linden Krister, Virpioja Sami 2005: *Morfessor and Hutmegs: Unsupervised Morpheme Segmentation for Highly-Inflecting and Compounding Languages*. Prieiga internete: https://www.researchgate.net/publication/228628569_Morfessor_and_hutmegs_Unsupervised_morpheme_segmentation_for_highly-inflecting_and_compounding_languages [žiūrėta 2022-11-22].

D

Dąbrowska Ewa 2015: What exactly is Universal Grammar, and has anyone seen it? – *Frontiers in Psychology* 6(852). Prieiga internete: https://www.researchgate.net/publication/279967160_What_exactly_is_Universal_Grammar_and_has_anyone_seen_it [žiūrėta 2022-11-22].

Dadurkevičius Virginijus 2017: Lietuvių kalbos morfologija atvirojo kodo Hunspell platformoje. – *Bendrinė kalba* 90, 1–15. Prieiga internete: http://www.bendrinekalba.lt/Straipsniai/90/Dadurkevicius_BK_90_straipsnis.pdf [žiūrėta 2022-11-22].

Dadurkevičius Virginijus, Petrauskaitė Rūta 2022: Corpus-Based Methods for Assessment of Traditional Dictionaries. – *Human Language Technologies – The Baltic Perspective A. Utka et al. (Eds.)*, 123–126. Prieiga internete: https://pdfs.semanticscholar.org/aa43/651066880ee574d3018363591453d7e721bd.pdf [žiūrėta 2022-11-22].

Dagienė Valentina, Grigas Gintautas 2007: *Programavimo kalbų teoriniai pagrindai*: mokymo priemonė bakalauro studijų programos „Matematikos ir informatikos mokymas“ studentams, Vilniaus universitetas. Prieiga internete: https://www.yumpu.com/lt/document/read/52424601/programavimo-kalba-teoriniai-pagrindai [žiūrėta 2022-11-22].

Daudaravičius Vidas 2012: *Teksto skaidymas pastoviųjų junginių segmentais*: daktaro disertacijos santrauka, Kaunas: Vytauto Didžiojo universitetas. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/124748/1/vidas_daudaravicius_dd.pdf [žiūrėta 2022-11-22].

Daudaravičius Vidas 2006: Pradžia į begalybę. – *Darbai ir dienos* 45, 7–18. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/32407/1/ISSN2335-8769_2006_N_45.PG_7-18.pdf [žiūrėta 2022-11-22].

DeRose Steven 1990: *Stochastic Methods for Resolution of Grammatical Category Ambiguity in Inflected and Uninflected Languages*: Ph.D. Dissertation, Providence, RI: Brown University Department of Cognitive and Linguistic Sciences. Prieiga internete: http://www.derose.net/steve/writings/dissertation/Diss.2.Foundations.html [žiūrėta 2022-11-22].

Donohue Mark 2011: *Grammar sketch outlines*. Prieiga internete: https://www.eva.mpg.de/lingua/tools-at-lingboard/pdf/donohue_grammar_sketches.pdf [žiūrėta 2022-11-22].

Dumčius Jonas, Kuzavinis Kazimieras, Mironas Ričardas 2010: *Elementa latīna*, Vilnius: Mokslo ir enciklopedijų leidybos centras.

Dumčius Jonas 2011: *Trumpa istorinė graikų kalbos gramatika*, Vilnius: Skaitmeninės filologijos centras. Prieiga internete: https://dokumen.tips/download/link/trumpa-istorin-graik-kalbos-gramatika [žiūrėta 2022-11-22].

E

Eastwood John 2002: *Oxford Guide to English Grammar*, Oxford: Oxford University Press. Prieiga internete: https://ia600305.us.archive.org/31/items/ilhem_20150408_1814/[John_Eastwood]_Oxford_Guide_to_English_Grammar.pdf [žiūrėta 2022-11-22].

Eigminas Kazimieras, Stundžia Bonifacas 1997: *Sapūno ir Šulco gramatika*, Vilnius: Mokslo ir enciklopedijų leidybos institutas.

Evans Nicholas, Levinson Stephen 2009: The myth of language universals: Language diversity and its importance for cognitive science. – *Behavioural and Brain Sciences* 32, 429–492. Doi: 10.1017/S0140525X0999094X. Prieiga internete: https://www.researchgate.net/publication/38036684_The_Myth_of_Language_Universals_Language_Diversity_and_Its_Importance_for_Cognitive_Science [žiūrėta 2022-11-22].

Everet Daniel 2005: Cultural Constraints on Grammar and Cognition in Piraha. – *Current Anthropology* 46(4), 621–646. Prieiga internete: https://www.researchgate.net/publication/215991936_Cultural_Constraints_on_Grammar_and_Cognition_in_Piraha_Another_Look_at_the_Design_Features_of_Human_Language [žiūrėta 2022-11-22].

F

Faruqui Manaal, McDonald Ryan, Soricut Radu 2016: Morpho-syntactic Lexicon Generation Using Graph-based Semi-supervised Learning. – *Transactions of the Association for Computational Linguistics* 4, 1–16. Prieiga internete: https://arxiv.org/pdf/1512.05030.pdf [žiūrėta 2022-11-22].

Floyd Simon 2011: Re-discovering the Quechua adjective. – *Linguistic Typology* 15, 25–63. Prieiga internete: https://www.researchgate.net/publication/228835457_Re-discovering_the_Quechua_adjective [žiūrėta 2022-11-22].

Fukushima Kunihiko 1980: Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. – *Biological cybernethics* 36, 193–202. Prieiga internete: https://www.cs.princeton.edu/courses/archive/spr08/cos598B/Readings/Fukushima1980.pdf [žiūrėta 2022-11-22].

G

Gaivenis Kazimieras, Keinys Stasys 1990: *Kalbotyros terminų žodynas*, Kaunas: Šviesa.

Geitgey Adam 2016: Language Translation with Deep Learning and the Magic of Sequences. – *Machine learning is fun* 5. Prieiga internete: https://medium.com/@ageitgey/machinelearning-is-fun-part-5-language-translation-with-deep-learning-and-the-magic-ofsequences-2ace0acca0aa [žiūrėta 2022-11-22].

Gelbukh Alexander, Grigori Sidorov 2003: Approach to Construction of Automatic Morphological Analysis Systems for Inflective Languages with Little Effort. – A. Gelbukh (ed.), *CICLing*, LNCS 2588, 215–220. Prieiga internete: https://www.researchgate.net/publication/221628854_Approach_to_Construction_of_Automatic_Morphological_Analysis_Systems_for_Inflective_Languages_with_Little_Effort [žiūrėta 2022-11-22].

Girdenis Aleksas, Žulys Vladas 1973: Lietuvių kalbos gramatika I. – *Baltistica* 9(2), 203–2014. Prieiga internete: http://www.baltistica.lt/index.php/baltistica/article/view/1832 [žiūrėta 2022-11-22].

Goldsmith John 2001: Abstract Unsupervised Learning of the Morphology of a Natural Language. – *Computational Linguistics* 27(2), 153–198. Prieiga internete: https://www.researchgate.net/publication/262282564_Abstract_Unsupervised_Learning_of_the_Morphology_of_a_Natural_Language [žiūrėta 2022-11-22].

Goldsmith John 2000: Linguistica: An Automatic Morphological Analyser. – To appear in John Boyle, Jung-Hyuck Lee, and Arika Okrent, *CLS 36 (Papers from the 36th Meeting of the Chicago Linguistics Society)* 1: *The Main Session*. Prieiga internete: https://www.researchgate.net/publication/246700009_Linguistica_An_Automatic_Morphological_Analyzer [žiūrėta 2022-11-22].

Greenbaum Sidney 1996: *The Oxford English Grammar*, Oxford: Oxford University Press.

Grumadienė Laima 2002: Dabartinės rašomosios lietuvių kalbos dažninis žodynas ir jo bazė. – *Acta Linguistica Lithuanica* XLVI, 19–37. Prieiga internete: http://etalpykla.lituanistikadb.lt/fedora/objects/LT-LDB-0001:J.04~2002~1367164490423/datastreams/DS.002.0.01.ARTIC/content [žiūrėta 2022-11-22].

Grumadienė Laima, Žilinskienė Vida 1998: *Dabartinės rašomosios lietuvių kalbos dažninis žodynas (abėcėlės tvarka)*, Vilnius: Matematikos ir informatikos institutas, Lietuvių kalbos institutas.

Grumadienė Laima, Žilinskienė Vida 1997: *Dažninis dabartinės rašomosios lietuvių kalbos žodynas*, Vilnius: Mokslo aidai.

Grūzītis Normunds, Dannélls Dana 2015: A Multilingual FrameNet-based Grammar and Lexicon for Controlled Natural Language. – *Language Resources and Evaluation* 51(1), 37–66. Prieiga internete: https://arxiv.org/pdf/1511.03924.pdf [žiūrėta 2022-11-22].

Guilbaud Jean-Philippe 1987: Principles and Results of a French MT System at Grenoble University (Geta). – S. Michaelson, Y. Wilks (series Ed.), *Machine Translation Today: The State of The Art. – Information Technology Series*, Edinburgh: Edinburgh University Press, 278–318.

Gulbinas Artūras 2019: *Giliųjų neuroninių tinklų taikymo kelio trūkių aptikimui nuotraukose tyrimas*: baigiamasis magistro projektas, Kauno technologijos universitetas. Prieiga internete: https://epubl.ktu.edu/object/elaba:37891115/ [žiūrėta 2022-11-22].

H

Hallgren Thomas, Enache Ramona, Ranta Aarne 2015: A Cloud-Based Editor for multilingual Grammars, Digital Grammar. – *The 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural language Processing: Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop*, Taberg: Taberg Media Group AB, 41–48. Prieiga internete: https://aclanthology.org/W15-3306.pdf [žiūrėta 2022-11-22].

Hanks Patrick 2004: Corpus pattern analysis. – G. Williams, S. Vessier (eds.), *Proceedings of the 11th EURALEX International Congress* 1, 2004 Jul 6–10, Lorient. France: Université de Bretagne-Sud, 87–97. Prieiga internete: https://www.researchgate.net/publication/228574527_Corpus_pattern_analysis [žiūrėta 2022-11-22].

Haspelmath Martin 2012: How to compare major word classes across the world's languages. – *UCLA Working Paper in Linguistics, Theories in Everything* 17(16), 109–130. Prieiga internete: http://phonetics.linguistics.ucla.edu/wpl/issues/wpl17/papers/16_haspelmath.pdf [žiūrėta 2022-11-22].

Helbig Gerhard, Buscha Joachim 1989: *Deutsche Grammatik*, Leipzig: VEB Verlag Enzyklopädie. Prieiga internete: https://kupdf.net/queue/helbig-buscha-deutsche-grammatik_58e14ed6dc0d60b2188970dc_pdf?queue_id=-1&x=1638976873&z=OTAuMTQwLjIzNC4yMTM= [žiūrėta 2022-11-22].

Hellwig Peter 2003: *DUG (Dependency Unification Grammar)*. Prieiga internete: https://www.cl.uni-heidelberg.de/~hellwig/dug-2003.pdf [žiūrėta 2007-06-18].

Hellwig Peter 2002: *Dependency Unification Grammar*. Prieiga internete: https://www.cl.uni-heidelberg.de/~hellwig/dug-2002.pdf [žiūrėta 2022-11-22].

Henisz-Dostert Bozena, Macdonald R. R., Zarechnak Michael 1979: Machine Translation. – *Trends in Linguistics. Studies and Monographs* 11, W. Winterk (ed.), New York: Mouton Publishers.

Holvoet Axel 2009: *Bendrosios sintaksės pagrindai*, Vilnius: Vilniaus universitetas ir asociacija „Academia Salensis“. Prieiga internete: http://www.academiasalensis.org/images/stories/bendrosios_sintakses_pagrindai.pdf [žiūrėta 2022-11-22].

Holvoet Axel, Mikulskas Rolandas 2009: Argumentų hierarchijos ir gramatinės funkcijos. – *Gramatinių funkcijų prigimtis ir raiška*, R. Mikulskas (red.), Vilnius: Academia Salensis. Prieiga internete: http://web.vu.lt/flf/b.stundzia/files/2019/05/Holvoet-2009-Difuziniai-subjektai-ir-objektai.pdf [žiūrėta 2022-11-22].

Hutchins John, Sommers Harold 1992: *An Introduction to Machine Translation*, London: Academic Press.

J

Jablonskis Jonas 1925: *Rygiškių Jono Lietuvių kalbos vadovėlis*, Kaunas: „Vaivos“ b-vė.

Jahn Elena 2020: *Ankündigung: Release 3 des Öffentlichen DGS-Korpus*. Prieiga internete: https://www.sign-lang.uni-hamburg.de/dgs-korpus/index.php/nachrichtenleser/ankuendigung-release-3-des-oeffentlichen-dgs-korpus.html [žiūrėta 2022-11-22].

Jensen Karen, Heidorn George, Richardson Stephen 1993: *Natural language processing: The PLNLP Approach*, Boston / London: Kluwer Academic Publishers.

Johnson Melvin, Schuster Mike, Le Quoc V., Krikun Maxim, Wu Yonghui, Chen Zhifeng, Thorat Nikhil, Viégas Fernanda, Wattenberg Martin, Corrado Greg, Hughes Macduff, Dean Jeffrey 2017. Google's Multilingual Neural Machine Translation System: Enabling Zero-Shot Translation. *Transactions of the Association for Computational Linguistics*, vol. 5, 339–351. Prieiga internete: https://aclanthology.org/Q17-1024.pdf [žiūrėta 2022-11-22].

Judžentis Artūras 2012: *Lietuvių kalbos gramatinės kategorijos*, Vilnius: Vilniaus universiteto leidykla.

Jung Walter 1967: *Grammatik der deutschen Sprache*, Leipzig: VEB Bibliographisches Institut.

K

Kairienė Audronė 2003: Dionisijas Trakietis. – *Visuotinė lietuvių enciklopedija* 4, Vilnius: Mokslo ir enciklopedijų leidybos institutas. Prieiga internete: https://www.vle.lt/straipsnis/dionisijas-trakietis/ [žiūrėta 2022-11-22].

Kay Martin, Gawron Mark, Norvig Peter 1994: *Verbmobil: A Translation System for Face-to-Face Dialog*, Stanford: CSLI.

Kapočiūtė-Dzikienė Jurgita, Damaševičius Robertas 2020: Coarse-Grained vs. Fine-Grained Lithuanian Dependency Parsing. – *Intelligent Algorithms in Software EngineeringCSOC 2020*, Cham: Springer, 450–464.

Kapočiūtė-Dzikienė Jurgita, Davidsonas Andrius, Vidugirienė Aušra 2017: Character-Based Machine Learning vs. Language Modeling for Diacritics Restoration. – *Information Technology and Control* 4(46), 508–520. ISSN 1392-124X. Prieiga internete: https://itc.ktu.lt/index.php/ITC/article/view/18066 [žiūrėta 2022-11-22].

Kapočiūtė-Dzikienė Jurgita, Rimkutė Erika, Boizou Loic 2017: A Comparizon of Lithuanian Morphological analyzers. – *Text, speech, and dialogue: 20th international conference proceedings*, K. Ekštein, V. Matoušek (eds.), Prague, Czech Republic, August 27–31, 47–56.

Keinys Stasys (vyr. red.), Klimavičius Jonas, Paulauskas Jonas, Pikčilingis Juozas, Sližienė Nijolė, Ulvydas Kazys, Vitkauskas Vytautas 1993: *Dabartinės lietuvių kalbos žodynas*, Vilnius: Mokslo ir enciklopedijų leidykla.

Kenny Doroti 2022: Human and machine translation. – *Machine translation for everyone / Empowering users in the age of artificial intelligence*, D. Kenny (ed.) 23–50. Prieiga internete: https://langsci-press.org/catalog/book/342 [žiūrėta 2022-11-22].

Kilgarriff Adam 2013: *Exploring Variation in Lexis and Genre in the Sketch Engine*. Prieiga internete: http://www.google.lt/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjv0uyy1e3sAhWMHHcKHVSPDyQQFjACegQIARAC&url=http%3A%2F%2Fkilgarriff.co.uk%2FPublications%2FKilgarriff_LSB.ppt%3Fformat%3Draw&usg=AOvVaw3Dp0Lfl124dBJN_mV0gu81 [žiūrėta 2022-11-22].

King Phil 2015: Cutting up Words: Morphology – Prefixes, Stems and Suffixes. – *Linguistics Modules for Language Discovery in the Pacific*, Pacific institute of Languages, Arts and Translation. Prieiga internete: https://www.youtube.com/watch?v=8ypQq5MvT24 [žiūrėta 2022-11-22].

Kniūkšta Pranas 2007: *Lietuvių kalbos žinynas*, Kaunas: „Šviesa“.

Köhler Reinhard 2012: *Quantitative Syntax Analysis / Quantitative Linguistics 65*, ed. by R. Köhler, G. Altmann, P. Grzybek, De G. Mouton.

Koskenniemi Kimmo 1983: Two-Level Model for Morphological Analysis. – *IJCAI 83 – Proceedings of the Tenth International Joint Conference* 2, 683–685. Prieiga internete: https://www.researchgate.net/publication/220813729_Two-Level_Model_for_Morphological_Analysis [žiūrėta 2022-11-22].

Kovalevskaitė Jolanta, Boizou Loic, Bielinskienė Agnė, Jancaitė Laima, Rimkutė Erika 2020: The First Corpus-Driven Lexical Database of Lithuanian as L2. – *Human Language Technologies – The Baltic Perspective*, A. Utka et al. (eds.), 245–252. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/110430/2/ISBN9781643681177.PG_245-252.pdf [žiūrėta 2022-11-22].

Kovalevskaitė Jolanta, Rimkutė Erika 2022: *Mokomasis lietuvių kalbos vartosenos leksikonas* – nauja tekstyno pagrindu parengta leksinė bazė. *Darnioji daugiakalbystė* 20, 154–193. Prieiga internete: https://sciendo.com/article/10.2478/sm-2022-0007 [žiūrėta 2022-11-22].

Kovalevskaitė Jolanta, Rimkutė Erika, Vilkaitė-Lozdienė Laura 2020: Light Verb Constructions in Lithuanian: Identification and Classification. – *Studies about Languages* 36, 5–16. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/109084/2/ISSN2029-7203_2020_N_36.PG_5-16.pdf [žiūrėta 2022-11-22].

Kuosienė Monika 1986: *Morfeminė analizė*, Šiauliai: K. Preikšo pedagoginis institutas.

Kuršaitis Frydrichas 2013: *Lietuvių kalbos gramatika 1876*, Vilnius: Vilniaus universiteto leidykla.

Kvietkauskas Valdemaras (ats. red.) 1985: *Tarptautinių žodžių žodynas*, Vilnius: Vyriausioji enciklopedijų redakcija.

L

Labutis Vitas 2002: *Lietuvių kalbos sintaksė*, Vilnius: Vilniaus universiteto leidykla.

Levane Kristine, Spektors Andrejs 2000: Morphemic Analysis and Morphological Tagging of Latvian Corpus. – *Proceedings of the Second International Conference on Language Resources and Evaluation* 2, Athens, Greece, May 31 – June 2, 1095–1098. Prieiga internete: http://www.lrec-conf.org/proceedings/lrec2000/pdf/107.pdf [žiūrėta 2022-11-22].

Liubinas Vilmantas 2021: CEF automatinio vertimo platforma. – *The third European Language Resource Coordination (ELRC) Workshop in Lithuania*. Prieiga internete: https://lr-coordination.eu/lt/lithuania3rd 2:13:40 [žiūrėta 2022-11-22].

M

Marcinkevičienė Rūta 2002: *Tekstynų lingvistika ir lietuvių kalbos vartosena*: habilitacinis darbas [rankraštis], Kaunas: Vytauto Didžiojo universitetas. Prieiga internete: https://etalpykla.lituanistikadb.lt/object/LT-LDB-0001:E.02~2002~1367158571385/E.02~2002~1367158571385.pdf [žiūrėta 2022-11-22].

Marcinkevičienė Rūta 2000: Tekstynų lingvistika. – *Darbai ir dienos* 24, 7–64. Prieiga internete: http://donelaitis.vdu.lt/publikacijos/marcinkeviciene.pdf [žiūrėta 2021-12-02].

Marcinkevičienė Rūta 1997: Tekstynų lingvistika ir lietuvių kalbos tekstynas. – *Lituanistica* 1(29), 58–78. Teksto prieiga internete: http://donelaitis.vdu.lt/publikacijos/lietka.htm [žiūrėta 2021-12-02].

Metuzale-Kangere Baiba 1985: *A Derivational Dictionary of Latvian / Latviešu Valodas Atvasinajumu Vardnica*, Hamburg: John Benjamins Pub Co.

McCrum Robert 2012: Daniel Everett: 'There is no such thing as universal grammar'. – *The Gardian*, The guardian News and Media Limited. Prieiga internete: https://www.theguardian.com/technology/2012/mar/25/daniel-everett-human-language-piraha [žiūrėta 2022-11-22].

McDonald Ryan, Nivre Joakim, Quirmbach-Brundage Yvonne, Goldberg Yoav, Das Dipanjan, Ganchev Kuzman, Hall Keith, Petrov Slav, Zhang Hao, Täckström Oscar, Bedini Claudia, Castelló Núria Bertomeu, Lee Jungmee 2013: Universal Dependency Annotation for Multilingual Parsing. – *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, Sofia, Bulgaria, August 4–9, 92–97. Prieiga internete: https://aclanthology.org/P13-2017.pdf [žiūrėta 2022-11-22].

McDonald Ryan, Pereira Fernando, Ribarov Kiril, Hajič Jan 2005: *Non-projective Dependency Parsing using Spanning Tree Algorithms*. Prieiga internete: https://ryanmcd.github.io/papers/nonprojectiveHLT-EMNLP2005.pdf [žiūrėta 2022-11-22].

McGilvray James 2018: Noam Chomsky. – *Encyclopaedia Britannica*, Encyclopaedia Britannica inc. Prieiga internete: https://www.britannica.com/biography/Noam-Chomsky#ref1033676 [žiūrėta 2022-11-22].

Mockus Darius 2018: Kalbos pinklės. – *Psytechnologijos*. Prieiga internete: http://www.psytechnologijos.lt/kalba/kalbos-pinkles/ [žiūrėta 2019-02-10].

Murmulaitytė Daiva 2012: Lietuvių kalbos morfemikos ir žodžių darybos tyrimų perspektyvos. – *Žmogus ir žodis* 1(14), 96–102. Prieiga internete: https://etalpykla.lituanistikadb.lt/fedora/objects/LT-LDB-0001:J.04~2012~1367186411825/datastreams/DS.002.0.01.ARTIC/content [žiūrėta 2022-11-22].

N

Naktinienė Gertrūda (vyr. red.), Paulauskas Jonas, Petrokienė Ritutė, Vitkauskas Vytautas, Zabarskaitė Jolanta 2008: *Lietuvių kalbos žodynas* I–XX (1941–2002): elektroninis variantas. Prieiga internete: www.lkz.lt [žiūrėta 2022-11-22].

Nau Nicole, Arkadiev Peter 2015: Towards a standard of glossing Baltic languages: The Salos Glossing Rules. – *Baltic Linguistics* 6, 195–241. Prieiga internete: https://www.researchgate.net/publication/287990735_Towards_a_standard_of_glossing_Baltic_languages_The_Salos_Glossing_Rules [žiūrėta 2022-11-22].

Nguyen Anh, Yosinski Jason, Clune Jeff 2015: Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images. – *Computer Vision and Pattern Recognition (CVPR '15)*, IEEE. Prieiga internete: https://www.cv-foundation.org/openaccess/content_cvpr_2015/papers/Nguyen_Deep_Neural_Networks_2015_CVPR_paper.pdf [žiūrėta 2022-11-22].

Nielsen Michael 2018: *Neural Networks and Deep Learning*. Prieiga internete: https://static.latexstudio.net/article/2018/0912/neuralnetworksanddeeplearning.pdf [žiūrėta 2022-11-22].

Nießen Sonja, Ney Hermann 2000: Improving SMT quality with morpho-syntactic analysis. – *COLING '00 Proceedings of the 18th conference on computational linguistics* 2, 1081–1085. Prieiga internete: https://www.researchgate.net/publication/221101548_Improving_SMT_quality_with_morpho-syntactic_analysis [žiūrėta 2022-11-22].

Nirenburg Sergei 1987: *Machine Translation: Theoretical and Methodological Issues*, London: Cambridge University Press.

Nivre Joakim 2014: Universal dependencies for Swedish. – *Abstracts of the Fifth Swedish Language Technology Conference*, Uppsala, Sweden, 13-14 November, 2014, 1-3. Prieiga internete: https://www2.lingfil.uu.se/SLTC2014/abstracts/sltc2014_submission_7.pdf [žiūrėta 2022-11-22].

Nivre Joakim, Hall Johan, Nilsson Jens, Chanev Atanas, Eryigit Gulsen, Kubler Sandra, Marinov Svetoslav, Marsi Erwin 2007: MaltParser: A language-independent system for data-driven dependency parsing. – *Natural Language Engineering* 13(2), 95–135. Prieiga internete: https://www.researchgate.net/publication/220597254_MaltParser_A_language-independent_system_for_data-driven_dependency_parsing/citations [žiūrėta 2022-11-22].

Nordquist Richard 2018: Universal Grammar (UG). – *ThoughtCo*, Dec. 7. Prieiga internete: https://www.thoughtco.com/universal-grammar-1692571 [žiūrėta 2022-11-22].

Nothman Joel, Murphy Tara, Curran James 2009: AnalysingWikipedia and Gold-Standard Corpora for NER Training. – *Proceedings of the 12th Conference of the European Chapter of the ACL*, Athens, Greece, 30 March – 3 April 2009, Association for Computational Linguistics, 612–620.

O

Osborne Tymothy, Gerdes Kim 2019: The status of function words in dependency grammar: A critique of Universal Dependencies (UD). – *Glossa: a journal of general linguistics* 4(1): 17, 1–28. Prieiga internete: https://doi.org/10.5334/gjgl.537 [žiūrėta 2022-11-22].

P

Paikens Peteris 2007: Lexicon-Based Morphological Analysis of Latvian Language. – *Proceedings of the 3rd Baltic Conference on Human Language Technologies*, Kaunas, 235–240. Prieiga internete: https://www.researchgate.net/publication/230800061_Lexiconbased_morphological_analysis_of_Latvian_language [žiūrėta 2022-11-22].

Paikens Peteris, Rituma Laura, Pretkalniņa Lauma 2013: Morphological Analysis with Limited Resources: Latvian example. – *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, Linköping Electronic Conference Proceedings #85, 267–277. Prieiga internete: https://www.researchgate.net/publication/260907832_Morphological_analysis_with_limited_resources_Latvian_example [žiūrėta 2022-11-22].

Payne Thomas Edward 2010: *A Sample Grammatical Sketch of English*. Prieiga internete: https://pages.uoregon.edu/tpayne/engram.htm [žiūrėta 2022-11-22].

Pakerys Jurgis 2014: Naujųjų skolinių duomenų bazės veiksmažodžių morfologija. – *Taikomoji kalbotyra* 3, 1–26. Doi: 10.15388/TK.2014.17480. Prieiga internete: https://www.zurnalai.vu.lt/taikomojikalbotyra/article/view/17480/16652 [žiūrėta 2022-11-22].

Palubinskienė Elena, Čepaitienė Giedrė 2008: *Lietuvių kalba*: vadovėlis VII klasei, antroji knyga, Kaunas: „Šviesa".

Paukert Herbert, Holböck Susanne 2017: *DEUGRA – eine Deutsche Grammatik*, Version 7.1. Eigenverlag http://www.paukert.at. Prieiga internete: http://www.paukert.at/sprachen/deugra.pdf [žiūrėta 2022-11-22].

Paulauskienė Aldona 2015: *Svarbesniosios XX a. lietuvių kalbos gramatikos*, Vilnius: „Gimtasis žodis".

Paulauskienė Aldona 1994: *Lietuvių kalbos morfologija. Paskaitos lituanistams*, Vilnius: Mokslo ir enciklopedijų leidykla.

Perez-Ortiz Juan Antonio, Forcada Mikel L., Sanchez-Martinez Felipe 2022: How neural machine translation works. – *Machine translation for everyone / Empowering users in the age of artificial intelligence*, D. Kenny (ed.) 141–164. Prieiga internete: https://langsci-press.org/catalog/book/342 [žiūrėta 2022-11-22].

Petrov Slav, Das Dipanjan, McDonald Ryan 2012: A Universal Part of Speech Tagset. – *Proceedings of the Eight International Conference on Language Resources and Evaluation – LREC'12*, European Language Resources Association (ELRA), 2089–2096. Prieiga internete: http://www.lrec-conf.org/proceedings/lrec2012/pdf/274_Paper.pdf [žiūrėta 2022-11-22].

Pinnis Mārcis, Krišlauks Rihards, Rikters Matīss 2019: Tilde's Machine Translation Systems for WMT 2019. – *Proceedings of the Fourth Conference on Machine Translation (WMT)* 2: Shared Task Papers (Day 1), Florence, Italy, August 1–2, 327–334. Prieiga internete: https://www.researchgate.net/publication/335715317_Tilde%27s_Machine_Translation_Systems_for_WMT_2019 [žiūrėta 2022-11-22].

Porter Martin 1980: An algorithm for suffix stripping. Computer laboratory, Corn Exchange Street, Cambridge. – Tomek Strzalkowski, *Robust Text Processing in Automated Information Retrieval*, Courant Institute of Mathematical Sciences, New York, 313–316. Prieiga internete: https://www.cs.toronto.edu/~frank/csc2501/Readings/R2_Porter/Porter-1980.pdf [žiūrėta 2022-11-22].

Pretkalniņa Lauma, Rituma Laura 2013: Statistical syntactic parsing for Latvian. – *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*, Linköping Electronic Conference Proceedings #85, 279–289. Prieiga internete: https://www.researchgate.net/publication/260907757_Statistical_syntactic_parsing_for_Latvian [žiūrėta 2022-11-22].

Przepiorkowsky Adam 2008: *Powerzchniowe przetwarzanie języka polskiego*, Akademicka Oficyna Wydawnicza EXIT.

Q

Quiles Carlos, López-Menchero Fernando 2011: *A Grammar of Modern Indo-European*, Indo-European Language Association. Prieiga internete: https://indo-european.info/indo-european-grammar.html [žiūrėta 2022-11-22].

Quirk Randolph, Greenbaum Sidney, Leech Geoffrey, Svartvik Jan 1992: *A Grammar of Contemporary English*, Singapore: Longman Group Ltd.

R

Ralys Danielius Algirdas 2017: Mašininis vertimas lietuvių kalbai. – *Bendrinė kalba* 90. Prieiga internete: http://journals.lki.lt/bendrinekalba/article/view/157/150 [žiūrėta 2022-11-22].

Ramachandran Aiswarya 2018: *NLP Guide: Identifying Part of Speech Tags using Conditional Random Fields. Analytics Vidhya*. Prieiga internete: https://medium.com/analyticsvidhya/pos-tagging-using-conditional-random-fields-92077e5eaa31 [žiūrėta 2022-11-22].

Rambow Owen, Creswell Cassandre, Szekely Rachel, Tauber Harriet, Walker Marilyn 2000: A dependency treebank for English. – *LREC'02*. Prieiga internete: http://www.lrec-conf.org/proceedings/lrec2002/pdf/325.pdf [žiūrėta 2022-11-22].

Ranta Aarne 2017: *Explainable Machine Translation*. Prieiga internete: http://www.grammaticalframework.org/~aarne/xmt-2017.pdf [žiūrėta 2022-11-22].

Ranta Aarne 2015: Data-Driven Documentation. – *A Technique for Reliable Multilingual Information Access*. Prieiga internete: http://www.grammaticalframework.org/~aarne/pic-2015-abstract.pdf [žiūrėta 2022-11-22].

Ranta Aarne 2014: *Embedded controlled languages*, *arXiv:1406.4057v1* [cs.CL]. Prieiga internete: https://arxiv.org/pdf/1406.4057.pdf [žiūrėta 2022-11-22].

Ranta Aarne 2013: *English: A Digital Grammar*. Prieiga internete: http://www.grammaticalframework.org/lib/doc/languages/gf-english.html [žiūrėta 2022-11-22].

Ranta Aarne 2011: *Grammatical Framework: Programming with Multilingual Grammars*, Stanford: CSLI Publications. Prieiga internete: http://www.grammaticalframework.org/gf-book/gf-book-slides.pdf [žiūrėta 2022-11-22].

Ranta Aarne 2009: The GF Resource Grammar Library. – *Linguistic Issues in Language Technology* 2(2), CSLI Publications. Prieiga internete: https://www.researchgate.net/publication/43647768_The_GF_Resource_Grammar_Library [žiūrėta 2022-11-22].

Rauh Gisa 2010: *Syntactic cateories. Teir Identificatios and Description in Linuistic Teories*, Oxford Linguistics, Oxford University Press. Prieiga internete: https://www.researchgate.net/publication/318034374_Gisa_Rauh_Syntactic_categories_Their_identification_and_description_in_linguistic_theories_Oxford_University_Press_Oxford_2010_xvii_436_pp [žiūrėta 2021-12-02].

Ren Xuancheng, Sun Xu, Wen Ji, Wei Bingshen, Zhan Weidong, Zhang Zhiyuan 2018: Building an Ellipsis-aware Chinese Dependency Treebank for Web Text. – *LREC Conference proceedings*. Prieiga internete: http://www.lrec-conf.org/proceedings/lrec2018/pdf/297.pdf [žiūrėta 2022-11-22].

Rieger Wilhelm 1903: *Zifferngrammatik welche mit Hilfe der Wörterbücher ein mechanisches Übersetzen aus einer Sprache in alle anderen ermöglicht*, Graz: Styria.

Rimkutė Erika 2017: *Dabartinės lietuvių kalbos žodžių morfeminė struktūra.* Prieiga internete: https://is.muni.cz/el/1421/podzim2017/BA480L/um/morfemika.pdf [žiūrėta 2022-11-22].

Rimkutė Erika 2006: Dabartinės lietuvių kalbos gramatinių formų vartosena morfologiškai anotuotame tekstyne. – *Lituanistika* 66(2), 34–55. Prieiga internete: https://etalpykla.lituanistikadb.lt/object/LT-LDB-0001:J.04~2006~1367156781132/J.04~2006~1367156781132.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Daudaravičius Vidas 2007: Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. – *Kalbų studijos* 11, 30–35. Prieiga internete: https://www.kalbos.lt/archyvas3.html [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011a: *Abėcėlinis lietuvių kalbos morfemikos žodynas* I, VDU: Kaunas. Prieiga internete: http://donelaitis.vdu.lt/lkk/pdf/AbcI.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011b: *Abėcėlinis lietuvių kalbos morfemikos žodynas* II, VDU: Kaunas. Prieiga internete: http://donelaitis.vdu.lt/lkk/pdf/AbcII.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011c: *Abėcėlinis lietuvių kalbos morfemikos žodynas* III, VDU: Kaunas. Prieiga internete: http://donelaitis.vdu.lt/lkk/pdf/AbcIII.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011d: *Dažninis lietuvių kalbos morfemikos žodynas* I, VDU: Kaunas. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/249/5/ISBN9789955126942.D_1.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011e: *Dažninis lietuvių kalbos morfemikos žodynas* II, VDU: Kaunas. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/249/6/ISBN9789955127369.D_2.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011f: *Dažninis lietuvių kalbos morfemikos žodynas* III, VDU: Kaunas. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/249/7/ISBN9789955127376.D_3.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011g: *Atgalinis lietuvių kalbos morfemikos žodynas* I, VDU: Kaunas. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/242/5/ISBN9789955126928.D_1.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011h: *Atgalinis lietuvių kalbos morfemikos žodynas* II, VDU: Kaunas. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/242/6/ISBN9789955127291.D_2.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2011j: *Atgalinis lietuvių kalbos morfemikos žodynas* III, VDU: Kaunas. Prieiga internete: https://www.vdu.lt/cris/bitstream/20.500.12259/242/7/ISBN9789955127307.D_3.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Raškinis Gailius 2010: Lietuvių kalbos veiksmažodžių morfeminė struktūra. – *Acta Linguistica Lithuanica* LXIV–LXV, 87–105. Prieiga internete: http://journals.lki.lt/actalinguisticalithuanica/article/view/1005/1095 [žiūrėta 2022-11-22].

Rimkutė Erika, Kazlauskienė Asta, Utka Andrius 2016: Morphemic Structure of Lithuanian Words. – *Open Linguistics* 2, 160–179. Prieiga internete: https://www.researchgate.net/publication/303908661_Morphemic_Structure_of_Lithuanian_Words [žiūrėta 2022-11-22].

Rimkutė Erika, Kovalevskaitė Jolanta 2008a: Mašininis vertimas: finišo tiesiosios nematyti... – *Gimtoji kalba* 5, 3–13. Prieiga internete: http://donelaitis.vdu.lt/lkk/pdf/MV.pdf [žiūrėta 2022-11-22].

Rimkutė Erika, Kovalevskaitė Jolanta 2008b: Tai, kas tinka žmogui, netinka mašininiam vertimui: žodynų problema. – *Gimtoji kalba* 2, 3–14. Prieiga internete: http://donelaitis.vdu.lt/lkk/pdf/zmog.pdf [žiūrėta 2022-11-22].

S

Salama Heba 2020: *Analysis of Egyptian child corpus based on automatic morphosyntactic tagging*: PhD Dissertation. Prieiga internete: https://www.researchgate.net/publication/344775132_Analysis_of_Egyptian_child_corpus_based_on_automatic_morphosyntactic_tagging [žiūrėta 2022-11-22].

Salama Heba, Alansary Sameh 2015: *Building a POS-Annotated Corpus For Egyptian Children*. Prieiga internete: https://www.researchgate.net/publication/299471250_Building_a_POS-Annotated_Corpus_For_Egyptian_Children [žiūrėta 2022-11-22].

Santorini Beatrice 1991: *Part-of-speech Tagging Guidelines for the Penn Treebank Project*. Prieiga internete: https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf [žiūrėta 2022-11-22].

Sasaki Felix, Witt Andreas 2004: Linguistische Korpora. – *Texttechnologie. Perspektiven und Anwendungen*, Tübingen: Stauffenburg, 195–216. Prieiga internete: https://core.ac.uk/download/pdf/83654078.pdf [žiūrėta 2022-11-22].

Schwanke Martina 1991: *Maschinelle Übersetzung: Ein Überblick über Theorie und Praxis*, Berlin: Springer-Verlag.

Sedlaček Radek 2004: *The Core of the Czech Derivational Dictionary*. Prieiga internete: http://www.lrec-conf.org/proceedings/lrec2004/pdf/696.pdf [žiūrėta 2022-11-22].

Silfverberg Miikka, Hulden Mans 2017: Automatic Morpheme Segmentation and Labeling in Universal Dependencies resources. – *Proceedings of the NoDaLiDa Workshop on Universal Dependencies*, 140–145. Prieiga internete: https://aclanthology.org/W17-0418.pdf [žiūrėta 2022-11-22].

Sinclair John 1992: The Automatic Analysis of Corpora. – *Directions in Corpus Linguistics*: *Proceedings of Nobel Symposium* 82, J. Startvik (ed.), Stockholm, 4–8 August 1991, Berlin: Walter de Gruyter, 379–397.

Skadiņš Raivis 2017: *Neural MT and other Language Technologies at TILDE*. Prieiga internete: http://school.grammaticalframework.org/2017/slides/Raivis-Neural_MT_at_Tilde.pdf [žiūrėta 2022-11-22].

Slavičkova Eleonora 2018: *Retrograde Morphemic Dictionary of Czech*, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Prieiga internete: http://hdl.handle.net/11234/1-2838 [žiūrėta 2022-11-22].

Sližienė Nijolė 1994–2005: *Lietuvių kalbos veiksmažodžių junglumo žodynas*, Vilnius: Lietuvių kalbos institutas.

Smit Petr, Virpioja Sami, Grönroos Stig-Arne, Kurimo Mikko 2014: Morfessor 2.0: Toolkit for statistical morphological segmentation. – *Conference: Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*. Prieiga internete: https://www.researchgate.net/publication/301404137_Morfessor_20_Toolkit_for_statistical_morphological_segmentation [žiūrėta 2022-11-22].

Stymne Sara 2014: *Transition-based dependency parsing*, Uppsala universitet: Department of Linguistics and Philology. Prieiga internete: https://cl.lingfil.uu.se/~sara/kurser/5LN455-2014/lectures/5LN455-F8.pdf [žiūrėta 2022-11-22].

Straka Milan, Strakova Jana 2017: Tokenizing, POS Tagging, Lemmatizing and Parsing UD 2.0 with UDPipe. *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, August 3-4, Vancouver, Canada, 88–99. Prieiga internete: https://aclanthology.org/K17-3.pdf [žiūrėta 2022-11-22].

Swamy M. N. S., Thulasiraman Krishnaiyan 1981: *Graphs, Networks, and Algorithms*, Wiley.

Szegedy Christian, Zaremba Wojciech, Sutskever Ilya, Bruna Joan, Erhan Dumitru, Goodfellow Ian, Fergus Rob 2014: *Intriguing properties of neural networks*. Prieiga internete: https://arxiv.org/pdf/1312.6199v4.pdf [žiūrėta 2022-11-22].

Š

Šveikauskienė Daiva 2021: Veiksmažodžio formų vartojamumas, išryškėjęs kuriant „Lietuvių kalbos gramatikos informacinę sistemą“. – *Bendrinė kalba* 94, 1–22. DOI: doi.org/10.35321/bkalba.2021.94.02. Prieiga internete: http://journals.lki.lt/bendrinekalba/article/view/2104/2206. [žiūrėta 2022-11-22].

Šveikauskienė Daiva 2018: Kokia kalbos dalis yra dalyvis? – *Bendrinė kalba* 91, 1–19. Prieiga internete: http://www.bendrinekalba.lt/?91 [žiūrėta 2022-11-22].

Šveikauskienė Daiva 2016: Lietuvių kalbos gramatikos informacinė sistema: I Morfologija. – *Lietuvių kalba* 10, 1–19. Prieiga internete: https://www.zurnalai.vu.lt/lietuviu-kalba/article/view/22601/21832 [žiūrėta 2022-11-22].

Šveikauskienė Daiva 2015a: Morphemic Structure of the Lithuanian Prefixes. – *Language: Meaning and Form / Language System and language use*, Conference proceedings, 189–197. Prieiga internete: https://www.lu.lv/fileadmin/user_upload/lu_portal/apgads/PDF/VNF_6/D_Sveikauskiene_VNF_6.pdf [žiūrėta 2022-11-22].

Šveikauskienė Daiva 2015b: High Quality Syntactic Annotated Corpus of Lithuanian – VILSINTEKS. – *Acta Linguistica Lithuanica* LXXIII, 252–267. Prieiga internete: http://lki.lt/wp-content/uploads/2017/06/acta_73_visas-ilovepdf-compressed.pdf [žiūrėta 2022-11-22].

Šveikauskienė Daiva 2014: Europos kalbų morfologinės analizės automatizavimas. – *Baltu filoloģija* XXIII(2), 101–116. Prieiga internete: https://dspace.lu.lv/dspace/bitstream/handle/7/31052/BaltuFilologija-23_2.pdf [žiūrėta 2022-11-22].

Šveikauskienė Daiva 2013: Lietuvių kalbos sintaksinė analizė. – *Lietuvių kalba* 7, 1–20. Prieiga internete: https://www.zurnalai.vu.lt/lietuviu-kalba/article/view/22685/21914 [žiūrėta 2022-11-22].

Šveikauskienė Daiva 2010: *Lietuvių kalbos vientisinių sakinių automatinė sintaksinė analizė*: daktaro disertacija, Vytauto Didžiojo universitetas.

Šveikauskienė Daiva, Ribikauskas Arūnas, Šveikauskas Vytautas 2019a. Lietuvių kalbos žodžių kompiuterinis kodavimas. – *The Word: Aspects of research* 23(1/2), 440–449. Prieiga internete: https://dom.lndb.lv/data/obj/836813.html [žiūrėta 2022-11-22].

Šveikauskienė Daiva, Ribikauskas Arūnas, Šveikauskas Vytautas 2017: Eine mehrsprachige Website zur Grammatik der litauischen Sprache. – *Acta Linguistica Lithuanica* LXXVII, 144–166. Prieiga internete: http://lki.lt/wp-content/uploads/2018/10/Acta_77_SP.pdf [žiūrėta 2022-11-22].

Šveikauskienė Daiva, Šveikauskas Vytautas 2019b: Lietuvių kalbos skaitmeninė gramatika. – *Lietuvių kalba* 13, 1–21. Prieiga internete: https://www.zurnalai.vu.lt/lietuviu-kalba/article/view/22488/21752 [žiūrėta 2022-11-22].

T

Tang Xuri 2006: English Morphological Analysis with Machine-learned Rules. – *PACLIC 20 Wuhan*, China, 1–3 November, 35–41. Prieiga internete: http://aclweb.org/anthology/Y/Y06/Y06-1005.pdf [žiūrėta 2022-11-22].

Taylor Ann, Marcus Mitchell, Santorini Beatrice 2003: The Penn Treebank: An Overview. – *Treebanks: Building and Using Parsed Corpora*, A. Abeille (ed.), Kluwer Academic Publishers, 5–22. Prieiga internete: https://www.researchgate.net/publication/2873803_The_Penn_Treebank_An_overview [žiūrėta 2022-11-22].

Tesnière Lucien 1959: *Éléments de syntaxe structurale*, Paris.

U

Ulvydas Kazys (red.) 1971: *Lietuvių kalbos gramatika: Morfologija* 2, Vilnius: „Mintis“.

Ulvydas Kazys (red.) 1965: *Lietuvių kalbos gramatika: Fonetika ir morfologija* 1, Vilnius: „Mintis“.

Utka Andrius, Amilevičius Darius, Krilavičius Tomas, Vitkutė-Adžgauskienė Daiva 2016: Overview of the Development of Language Resources and Technologies in Lithuania (2012–2015). – *Human Language Technologies – The Baltic Perspective*, I. Skadiņa and R. Rozis (eds.), 13–19. Prieiga internete: https://www.researchgate.net/publication/329209846_Overview_of_the_Development_of_Language_Resources_and_Technologies_in_Lithuania_2012-2015 [žiūrėta 2022-11-22].

V

Vaičiulytė-Semėnienė Loreta, Čižik-Prokaševa Veslava, Gritėnienė Aurelija, Liutkevičienė Danutė, Gaidienė Anželika 2022: *Lietuvio kaimynai: draugai ar priešai? Kalbinė vaizdinių analizė*. Kolektyvinė monografija, Vilnius: Lietuvių kalbos institutas. ISBN 978-609-411-328-4.

Vaitkevičiūtė Valerija 2007: *Tarptautinių žodžių žodynas*, Vilnius: Žodynas.

W

Wallis Sean, Nelson Gerald 2001: Knowledge Discovery in Grammatically Analysed Corpora. – *Data Mining and Knowledge Discovery* 5, 305–335. Prieiga internete: https://ccc.inaoep.mx/~villasen/bib/Knowledge%20Discovery%20in%20Grammatically%20Analysed%20Corpora.pdf [žiūrėta 2022-11-22].

Winograd Terry 1983: *Language as a Cognitive Process* I: *Syntax*, London: Addison-Wesley Publishing Company.

Wong Meng Weng 2017: *Open Problems in Computational Law*. Prieiga internete: https://docs.google.com/presentation/d/15FS3FwlILBVFRIhwQT3Szq4zjz208Hce3K5jf8Q-3o0/edit#slide=id.g228038fb28_0_3 [žiūrėta 2022-11-22].

Z

Zaikauskas Egidijus 2019: eTranslation viršūnės ir gelmės. – *Antrasis ELRC seminaras Lietuvoje*, Vilnius, 2019 m. vasario 1 d.

Zinkevičius Vytautas 2000: Lemuoklis – morfologinei analizei. – *Darbai ir dienos* 24, 245–273. Prieiga internete: http://donelaitis.vdu.lt/publikacijos/zinkevicius.pdf [žiūrėta 2021-12-02].

Ž

Žaliauskas Nikodemas 2017: *Kalbėtojo atpažinimas naudojantis dirbtiniais neuroniniais tinklais*: baigiamasis bakalauro darbas, Vilniaus universitetas. Prieiga internete: http://talpykla.elaba.lt/elaba-fedora/objects/elaba:23159558/datastreams/MAIN/content [žiūrėta 2022-11-22].

Žiugžda Juozas 1971: *Lietuvių kalbos gramatika* 1: *Fonetika ir morfologija*, Kaunas: „Šviesa".

Б

Бабайцева Вера Васильевна, Чеснокова Лилия Дмитриевна 2012: *Русский язык – Теория*, Москва: «Дрофа». Prieiga internete: https://kstu.kg/fileadmin/user_upload/russkii_jazyk.pdf [žiūrėta 2022-11-22].

Белецкая Ирина Платоновна 1983: Деревья зависимостей как инструмент стнтаксического анализа текста. – *Международный семинар по машинному переводу*, Москва: ВЦП.

Г

Грязнухина Татьяна Александровна 1999: *Синтаксический анализ научного текста на ЭВМ*, Киев: «Наукова думка».

Ш

Шведова Наталия Юльевна 1980: *Русская грамматика* 1, Москва: «Наука». Prieiga internete: http://rkiff.philol.msu.ru/wp-content/uploads/2020/05/%D0%A0%D1%83%D1%81%D1%81%D0%BA%D0%B0%D1%8F-%D0%B3%D1%80%D0%B0%D0%BC%D0%BC%D0%B0%D1%82%D0%B8%D0%BA%D0%B0.-%D0%A8%D0%B2%D0%B5%D0%B4%D0%BE%D0%B2%D0%B0.-%D0%A2%D0%BE%D0%BC-1.pdf [žiūrėta 2022-11-22].

ZUSAMMENFASSUNG

COMPUTERISIERUNG VON GRAMMATIK DER LITAUISCHEN SPRACHE

Nach dem Aufkommen von Computern begannen sie in alle Gebiete des Lebens einzudringen. Sprachen waren keine Ausnahme. Es ist schon viel in der Welt geleistet. In Litauen wurden auch mehrere Arbeiten durchgeführt. Im vorliegenden Text werden die Werke besprochen, in denen die Computerisierung der Grammatik unternommen wurde: Annotation der Textkorpora, morphologische Analysatoren sowie Parser, digitale Grammatik und Informationssystem für Grammatik der litauischen Sprache. Jedem Thema wird ein Passus gewidmet.

Schon im Mittelalter wurden die ersten Vorschläge gemacht, die Sprachen mit Zahlen zu verbinden. Und erst nach dem Erscheinen von Computern wurden diese Ideen verwirklicht. Die neuesten Technologien – neurale Netzwerke – waren erfolgreich und haben zu guten Leistungen in manchen Gebieten der Sprachverarbeitung geführt. Doch die 100-prozentige Genauigkeit ist noch nicht erreicht.

Am Anfang der Annotation der Korpora wurde es klar, dass die Vielfalt der Sprachen eine große Auswirkung für die Tags hat. Deshalb gibt es bis jetzt kein allgemeines Tagset. Bei der Annotation der litauischen Korpora werden auch die litauischen Tags neben der englischen aufgeführt.

Für die syntaktische Annotation der Sätze waren einige Formaten vorbereitet. Nicht alle aber sind gleich verbreitet. Am meisten ist wohl der PML benutzt, der auch für die litauische Sprache gebraucht wird.

Der an der Universität Vytautas des Großen vorbereitete morphologische Analysator benutzt die Regeln-basierte Methode. Er wurde auf der Plattform *Hunspell* geschaffen. Aufgrund statistischer Methoden wird die morphologische Analyse der litauischen Sprache mit *UDPipe* durchgeführt.

Die erste Datenbank für Morphemik, die am Institut für Mathematik und Informatik geschaffen wurde, umfasst die ausführliche Information über die

Morpheme des Wortes, einschließlich des Morphemtyps. Leider ist sie nicht frei zugänglich. Die im Internet veröffentlichte Datenbank für Morphemik (Universität Vytautas des Großen) führt die Wörter auf, die durch Bindestriche in Morpheme geteilt sind. Es gibt aber keine Information über den Morphemtyp.

Der Regeln-basierte syntaktische Analysator der Litauischen Sprache wurde auf der Website der Universität Vytautas des Großen bis Februar des Jahres 2020 zugänglich. Derzeit steht keine erneuerte Version zur Verfügung.

Der Modul der syntaktischen Analyse der litauischen Sprache von *UDPipe* funktioniert aufgrund der statistischen Methode. Für maschinelles Lernen wurde der syntaktisch annotierte Korpus der litauischen Sprache *ALKSNIS* benutzt. Die Fehler entstehen meistens in den Fällen, wenn litauische Sätze nur der litauischen Sprache eigene Merkmale besitzen, z. B. für die englische Sprache untypische Wortfolge u. a.

Es hat sich erwiesen, dass die Methoden, die für die englische Sprache gute Leistungen geben, nicht immer für andere Sprachen ebenso erfolgreich benutzt werden können.

Zurzeit ist nur eine sehr kurze Skizze der Grammatik für die litauische Sprache vorbereitet. Das Probebeispiel der digitalen Grammatik ist geschaffen.

Das Ziel des Informationssystems für Grammatik der litauischen Sprache ist die Vorbereitung der Sprachdokumentation.

Im Informationssystem gespeicherte Daten werden zweierlei dargestellt. Für den breiten Benutzerkreis werden die Daten möglichst einfach und verständlich geschildert. Für die wissenschaftlichen Forschungen wird das computerfreundliche Format benutzt.

Auf der Webseite ist jeweils sowohl morphologische, als auch morphemische Information über das Wort vorhanden. Der Typ des Morphems wird auch angezeigt und die Wortstruktur erläutert: Das Lemma, Grund- und Bestimmungswort für Ableitungen und Zusammensetzungen werden angegeben. Alle übrigen Formen eines Wortes können über den Links ÜBRIGE FORMEN abgerufen werden. Die Webseite wird in 7 Sprachen: Litauisch, Englisch, Deutsch, Französisch, Italienisch, Russisch und Japanisch geführt.

Vorläufig ist nur der morphologische Teil der Grammatik entwickelt. In der Zukunft hat man vor, auch die syntaktischen Daten in das Informationssystem einzutragen.

Während der Entwicklung des Informationssystems wurden die neuen Erscheinungen in der litauischen Grammatik bemerkt, die früher nie von Sprachwissenschaftlern beschrieben worden waren, z. B., das Fehlen mancher Formen im Paradigma eines bestimmten Wortes. Diese Erscheinung ist durch die Semantik bestimmt. Genauer gesagt, sie verursacht der Zwiespalt zwischen der lexikalischen und grammatischen Bedeutung des Wortes. Der Singular ist für die Verben mit der Bedeutung der Gruppenaktion unmöglich. Die von intransitiven Verben abgeleiteten Partizipien im Passiv können nur die Formen des Neutrums haben usw.

SUMMARY

THE COMPUTERISATION OF THE LITHUANIAN GRAMMAR

As soon as they made their first appearance, computers began to spread across most of the fields of human life. Languages are no exception here. A lot has been accomplished in the world. Many things have been accomplished in Lithuania, too. This study offers a description of endeavours in the field of computerisation of grammar, including corpus annotations, morphological analysers and parsers, digital grammar, and a grammar information system. Each individual chapter covers a particular subject.

The first efforts to combine languages with digits were made back in the Middle Ages. However, it was with the advent of the computer that these ideas started seeing some potential for implementation. The latest technology – neural networks – has produced decent results in some areas, yet 100 percent accuracy is still out of reach.

After they had started making corpus annotations, researchers have discovered that tags are highly affected by the diversity of languages – that is why no uniform annotation tag set has been developed yet. In addition to English tags, the morphologically annotated corpus of the Lithuanian language also provides Lithuanian tags.

Many different formats have been developed for the purposes of syntactic annotation, yet not all of them have followed a similar spread pattern. The Prague Markup Language, or PML designed by Prague researchers was probably the one that enjoyed the highest degree of popularity. It is also used to annotate sentences written in the Lithuanian language.

The morphological analyser developed by Vytautas Magnus University (VMU) on the basis of the *Hunspell* platform operates under a rule-based approach. Morphological analysis of the Lithuanian language grounded on statistical methods is performed by the *UDPipe* Lithuanian language module.

Developed by the Institute of Mathematics and Informatics, the first morphemic database contains detailed information about morphemes, including their types, yet its contents are not freely accessible. The VMU online morphemic database produces words broken down in hyphenated morphemes yet contains zero information about the type of the morpheme.

The rule-based parser had been accessible on VMU's website until February 2020 and not updated version of it has been made available as yet.

The *UDPipe* module is a parser of the Lithuanian language that uses statistic methods to function. The syntactically annotated Lithuanian corpus ALKSNIS was used for the machine learning. Any inaccuracies in parsing are primarily caused by sentences that carry specific qualities of the Lithuanian language, such as a peculiar ordering of words that cannot be typically found in the English language, and so on.

It appears that methods that successfully apply to the English language cannot always be used with other languages.

For now, only a pilot sample of a digital grammar of the Lithuanian language and a limited Sketch grammar are available.

The purpose of developing an information system for the grammar of the Lithuanian language is to draw documentation on the grammar of the Lithuanian language. The system stores two types of information: data designed for the wide public, which are available in a popular and comprehensive form, and a computer friendly format used for scientific research purposes. The website contains both morphological and morphemic data with indication of morpheme type. The structure of the word – the lemma and underlying words for derivatives – is reflected as well. All inflexional forms can be viewed by clicking the OTHER FORMS button. The information on the website is available in seven languages: Lithuanian, English, German, French, Italian, Russian, and Japanese. Only the model for the morphological segment is available at this time, with the syntactic segment slated for development some time in the future.

The development of the *Lithuanian Grammar Information System* (LIGIS) has highlighted new phenomena that have never been covered by linguists before: sometimes words do not have all of the paradigmatic forms, which is the product of the semantics of the word, or rather the difference between its semantic meaning and its grammatical meaning, for instance: verbs that denote a group action cannot have a

singular form; passive-voice participles made from intransitive verbs can only have neuter forms, and so on.

PRIEDAI

1 PRIEDAS: Paveikslėlių sąrašas

1. ĮVADAS

2. ANOTUOTI TEKSTYNAI

3. MORFOLOGIJOS KOMPIUTERIZAVIMAS

4. SINTAKSĖS KOMPIUTERIZAVIMAS

5. SKAITMENINĖ GRAMATIKA

6. LIETUVIŲ KALBOS GRAMATIKOS INFORMACINĖ SISTEMA

2 PRIEDAS: MATO failo PUB-014 fragmentas (TAB-WPL formatu)

<s>

Petras	Petras	Npmsnnf
Dabrišius	Dabrišius	Npmsnns

</s>

<p/>

<s>

ŽIEMA	žiema	Ncfsnn-
SU	su	Sgi
VILKAIS	vilkas	Ncmpin-

</s>

<p/>

<s>

Tik	tik	Rgp
temstant	temti	Vgap----n--n--
pasiekėme	pasiekti	Vgma1p--n--ni-
uolas	uola	Ncfpan-

<g/>

,	,	Tc
kurios	kuris	Pgfpnn
stovėjo	stovėti	Vgma3---n--ni-
kaip	kaip	Cg
vartai	vartai	Ncmpnn-
į	į	Sga
tarpeklį	tarpeklis	Ncmsan-

<g/>

.	.	Tp

</s>

<p/>

<s>

Vairuotojas	vairuotojas	Ncmsnn-
toliau	toli	Rgc
važiuoti	važiuoti	Vgi-----n--n--
atsisakė	atsisakyti	Vgma3---n--yi-

<g/>

.	.	Tp

</s>

<s>

Matyt	matyt	Qg

<g/>

,	,	Tc
jį	jis	Pgmsan
baugino	bauginti	Vgma3---n--ni-
stačios	status	Agpfpnn
akmeninės	akmeninis	Agpfpnn
sienos	siena	Ncfpnn-

<g/>

,	,	Tc
grėsmingai	grėsmingai	Rgp
kabančios	kaboti	Vgpp-pfannnn-p
virš	virš	Sgg
galvų	galva	Ncfpgn-

<g/>

.	.	Tp

</s>

3 PRIEDAS: MATO failo PUB-014 fragmentas (CoNLL-U formatu)

```
# newdoc id = pub-014-doc1
# newpar id = pub-014-doc1-p1
# sent_id = pub-014-doc1-s1
# text = Petras Dabrišius
1	Petras	Petras	PROPN	dkt.tikr.vyr.vns.V.	Case=Nom|Gender=Masc|Number=Sing	_	_	_	Multext=Npmsnnf
2	Dabrišius	Dabrišius	PROPN	dkt.tikr.vyr.vns.V.	Case=Nom|Gender=Masc|Number=Sing	_	_	_	Multext=Npmsnns

# newpar id = pub-014-doc1-p2
# sent_id = pub-014-doc1-s2
# text = ŽIEMA SU VILKAIS
1	ŽIEMA	žiema	NOUN	dkt.mot.vns.V.	Case=Nom|Gender=Fem|Number=Sing	_	_	_	Multext=Ncfsnn-
2	SU	su	ADP	prl.Įn.	Case=Ins	_	_	_	Multext=Sgi
3	VILKAIS	vilkas	NOUN	dkt.vyr.dgs.Įn.	Case=Ins|Gender=Masc|Number=Plur	_	_	_	Multext=Ncmpin-

# newpar id = pub-014-doc1-p3
# sent_id = pub-014-doc1-s3
# text = Tik temstant pasiekėme uolas, kurios stovėjo kaip vartai į tarpeklį.
```

1	Tik	tik	ADV	prv.nelygin.	Degree=Pos	_	_	_	Multext=Rgp											
2	temstant	temti	VERB	vksm.pad.es.	Polarity=Pos|Tense=Pres|VerbForm=Ger	_	_	_	Multext=Vgap----n--n--											
3	pasiekėme	pasiekti	VERB	vksm.asm.būt-k.1.dgs.tiesiog.	Mood=Ind|Number=Plur|Person=1|Polarity=Pos|Tense=Past|VerbForm=Fin	_	_	_	Multext=Vgma1p--n--ni-											
4	uolas	uola	NOUN	dkt.mot.dgs.G.	Case=Acc|Gender=Fem|Number=Plur	_	_	_	SpaceAfter=No|Multext=Ncfpan-											
5	,	,	PUNCT	skyr.	_	_	_	_	Multext=Tc											
6	kurios	kuris	PRON	įv.mot.dgs.V.	Case=Nom|Definite=Ind|Gender=Fem|Number=Plur	_	_	_	Multext=Pgfpnn											
7	stovėjo	stovėti	VERB	vksm.asm.būt-k.3.tiesiog.	Mood=Ind|Person=3|Polarity=Pos|Tense=Past|VerbForm=Fin	_	_	_	Multext=Vgma3---n--ni-											
8	kaip	kaip	SCONJ	jng.	_	_	_	_	Multext=Cg											
9	vartai	vartai	NOUN	dkt.vyr.dgs.V.	Case=Nom|Gender=Masc|Number=Plur	_	_	_	Multext=Ncmpnn-											
10	į	į	ADP	prl.G.	Case=Acc	_	_	_	Multext=Sga											
11	tarpeklį	tarpeklis	NOUN	dkt.vyr.vns.G.	Case=Acc|Gender=Masc|Number=Sing	_	_	_	SpaceAfter=No|Multext=Ncmsan-											
12	.	.	PUNCT	skyr.	_	_	_	_	Multext=Tp											

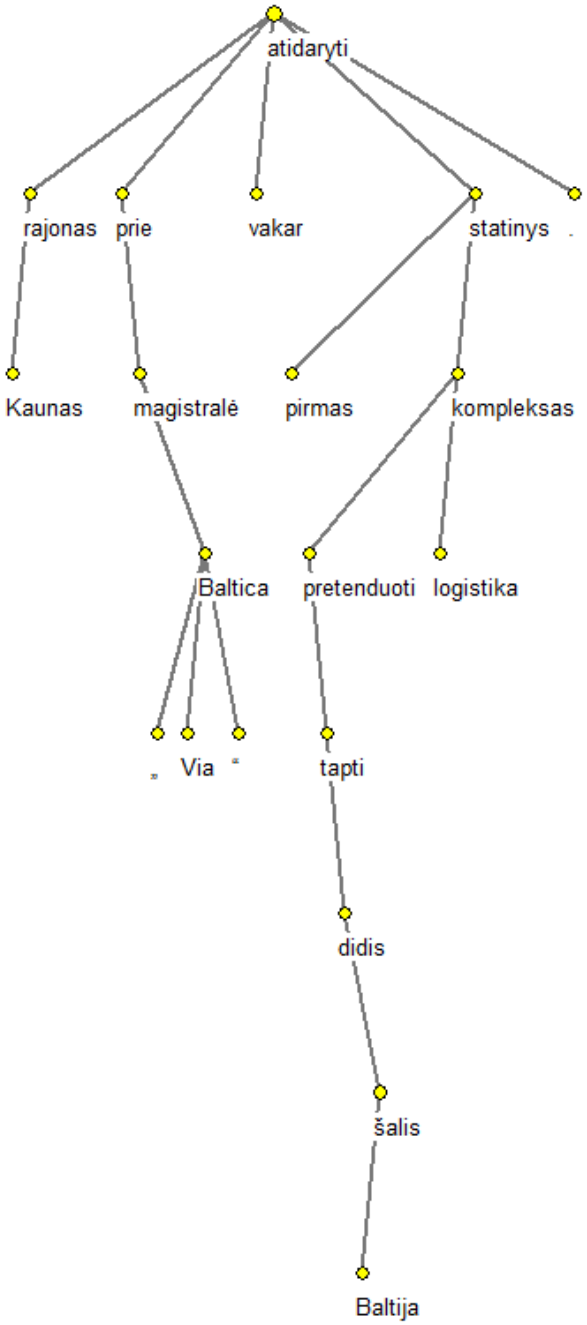
newpar id = pub-014-doc1-p4

sent_id = pub-014-doc1-s4

text = Vairuotojas toliau važiuoti atsisakė.

1	Vairuotojas	vairuotojas	NOUN	dkt.vyr.vns.V.	Case=Nom|Gender=Masc|Number=Sing	_	_	_	Multext=Ncmsnn-											
2	toliau	toli	ADV	prv.aukšt.	Degree=Cmp	_	_	_	Multext=Rgc											
3	važiuoti	važiuoti	VERB	vksm.bndr.	Polarity=Pos|VerbForm=Inf	_	_	_	Multext=Vgi-----n--n--											
4	atsisakė	atsisakyti	VERB	vksm.asm.būt-k.3.tiesiog.sngr.	Mood=Ind|Person=3|Polarity=Pos|Reflex=Yes|Tense=Past|VerbForm=Fin	_	_	_	SpaceAfter=No|Multext=Vgma3---n--yi-											
5	.	.	PUNCT	skyr.	_	_	_	_	Multext=Tp											

4 PRIEDAS: ALKSNIS 3.0 versijos sakinys, anotuotas PML formatu



PML formatu anotuotas sakinys *Kauno rajone prie magistralės „Via Baltica“ vakar atidarytas pirmasis pretenduojančio tapti didžiausiu Baltijos šalyse logistikos komplekso statinys.*

5 PRIEDAS: ALKSNIS 3.0 versijos sakinys, anotuotas CoNLL-U formatu

 # newdoc id = kd1-2

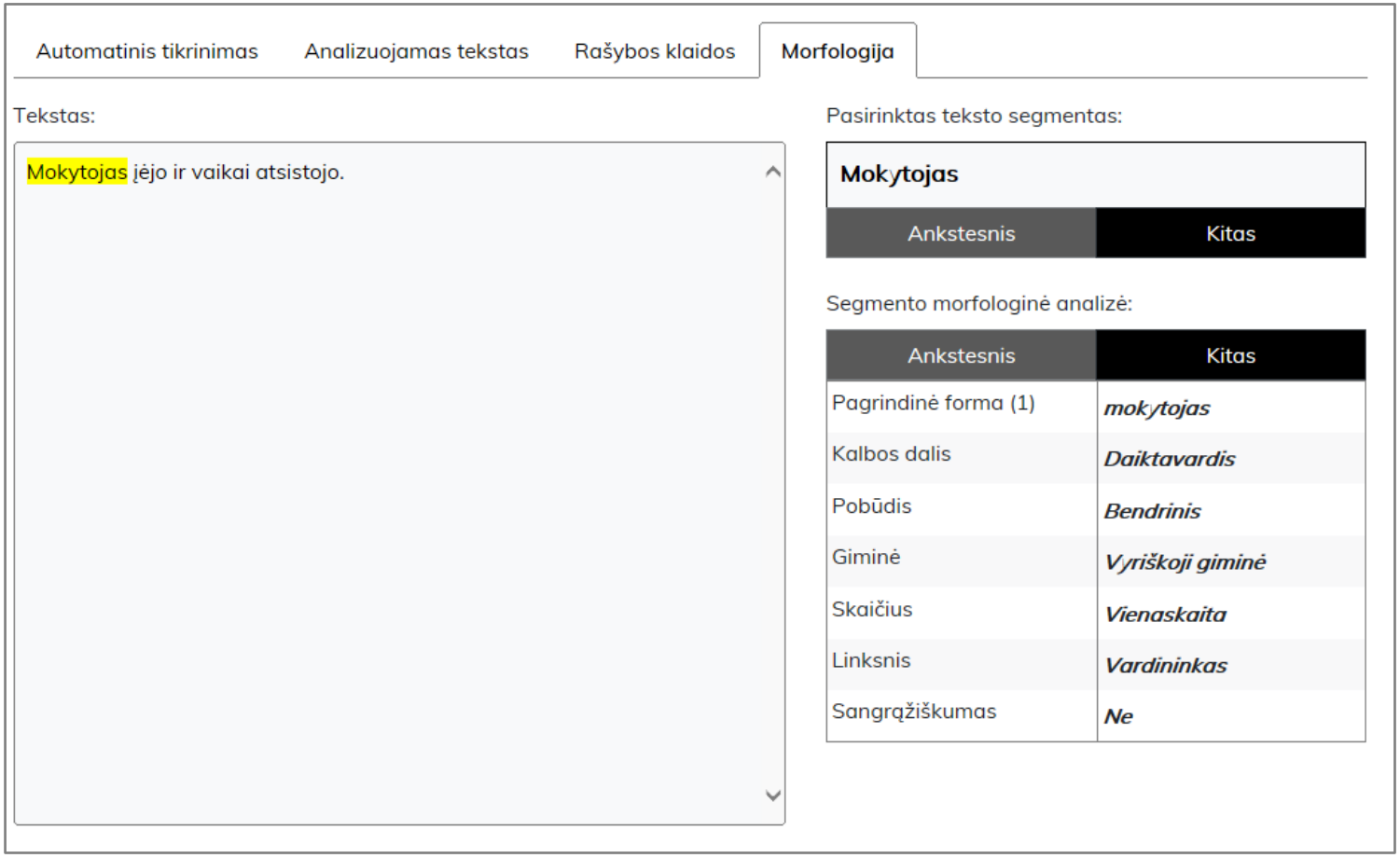
 # sent_id = kd1-2-s1

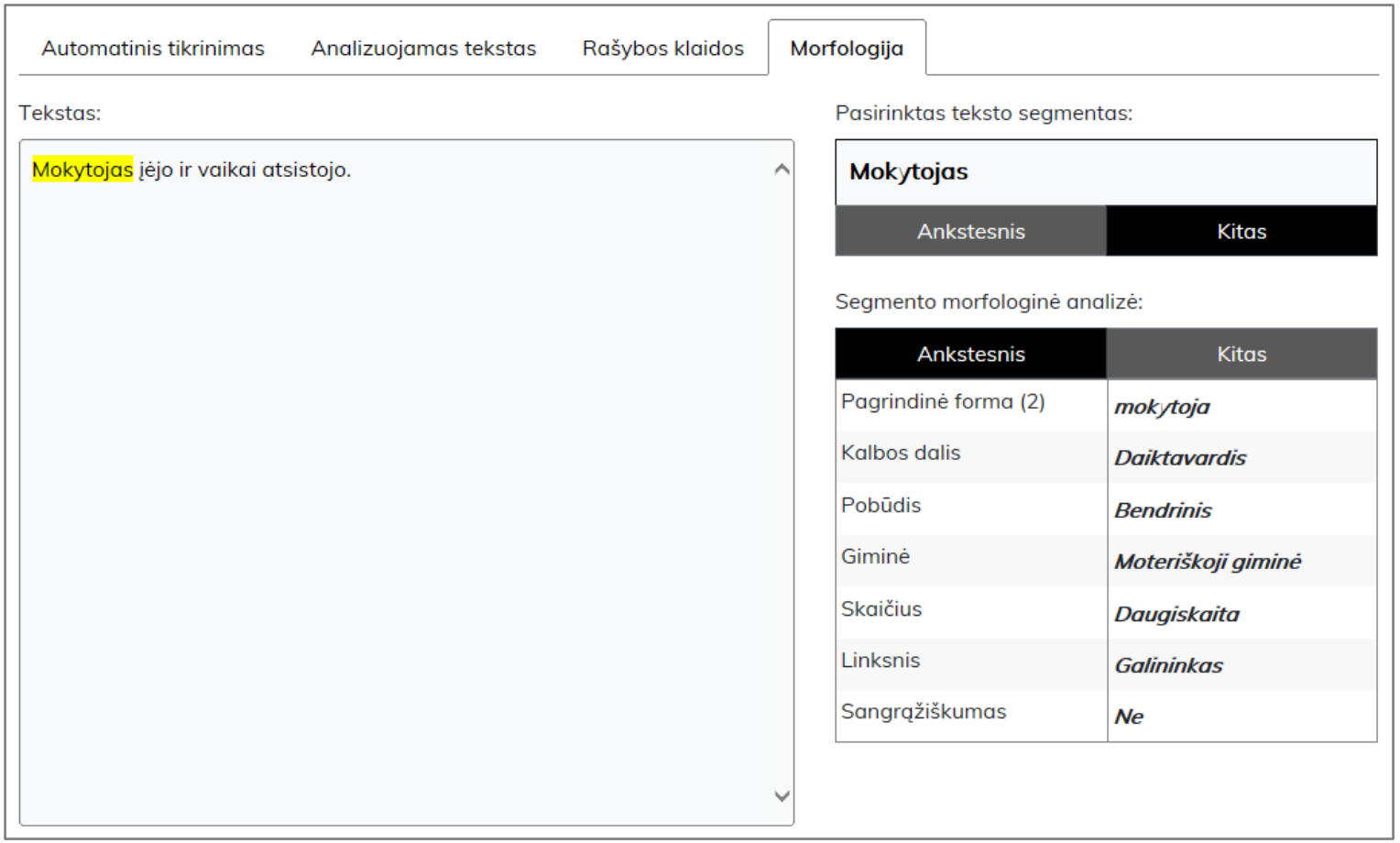
 # text = Kauno rajone prie magistralės „ Via Baltica “ vakar atidarytas pirmasis pretenduojančio tapti didžiausiu Baltijos šalyse logistikos komplekso statinys.

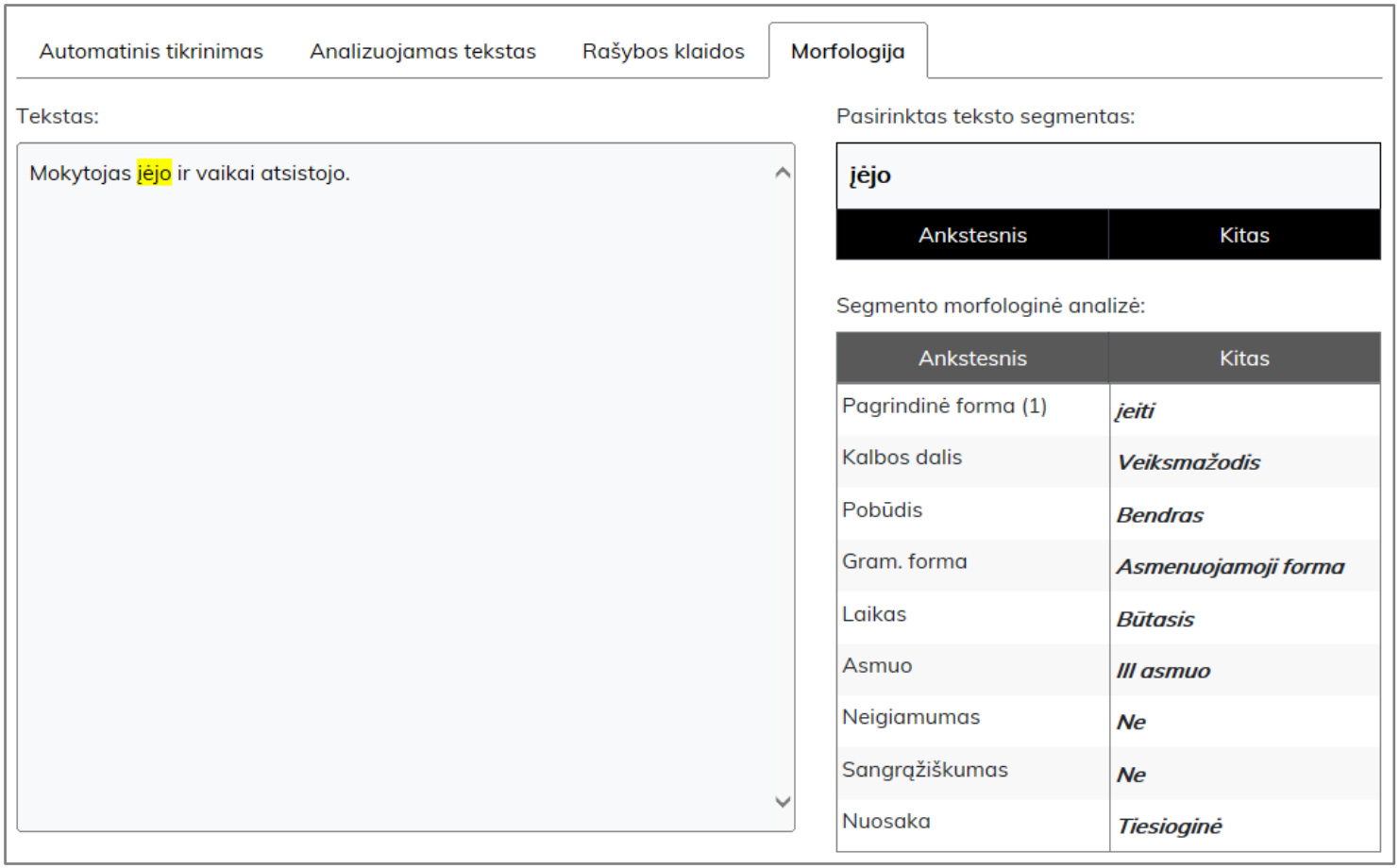
1	*Kauno*	*Kaunas*	*NOUN*	*dkt.tikr.vyr.vns.K.*		*Case=Gen|Gender=Masc|Number=Sing*	*2*	*Atr*	*2:Atr*	_		
2	*rajone*	*rajonas*	*NOUN*	*dkt.vyr.vns.Vt.*		*Case=Loc|Gender=Masc|Number=Sing*	*10*	*Adv*	*10:Adv*	_		
3	*prie*	*prie*	*ADP*	*prl.K.*	*Case=Gen*	*10*	*AuxP*	*10:AuxP*	_			
4	*magistralės*	*magistralė*	*NOUN*	*dkt.mot.vns.K.*	*Case=Gen|Gender=Fem|Number=Sing*		*3*	*Adv*	*3:Adv*	_		
5	*„*	*„*	*PUNCT*	*skyr.*	_	*7*	*Aux*	*7:Aux*	*SpaceAfter=No*			
6	*Via*	*Via*	*X*	*užs.*	*Foreign=Yes*	*7*	*AuxL*	*7:AuxL*	_			
7	*Baltica*	*Baltica*	*X*	*užs.*	*Foreign=Yes*	*4*	*Atr*	*4:Atr*	*SpaceAfter=No*			
8	*“*	*“*	*PUNCT*	*skyr.*	_	*7*	*Aux*	*7:Aux*	_			
9	*vakar*	*vakar*	*ADV*	*prv.nelygin.*	*Degree=Pos*	*10*	*Adv*	*10:Adv*	_			
10	*atidarytas*	*atidaryti*	*VERB*	*vksm.dlv.neveik.būt.vyr.vns.V.*			*Case=Nom|Gender=Masc|Number=Sing|Tense=Past|VerbForm=Part|Voice=Pass*	*0*	*PredN*	*0:PredN*	_	
11	*pirmasis*	*pirmas*	*ADJ*	*sktv.raid.kelint.įvardž.vyr.vns.V.*			*Case=Nom|Definite=Def|Gender=Masc|NumType=Ord|Number=Sing*	*19*	*Atr*	*19:Atr*	_	

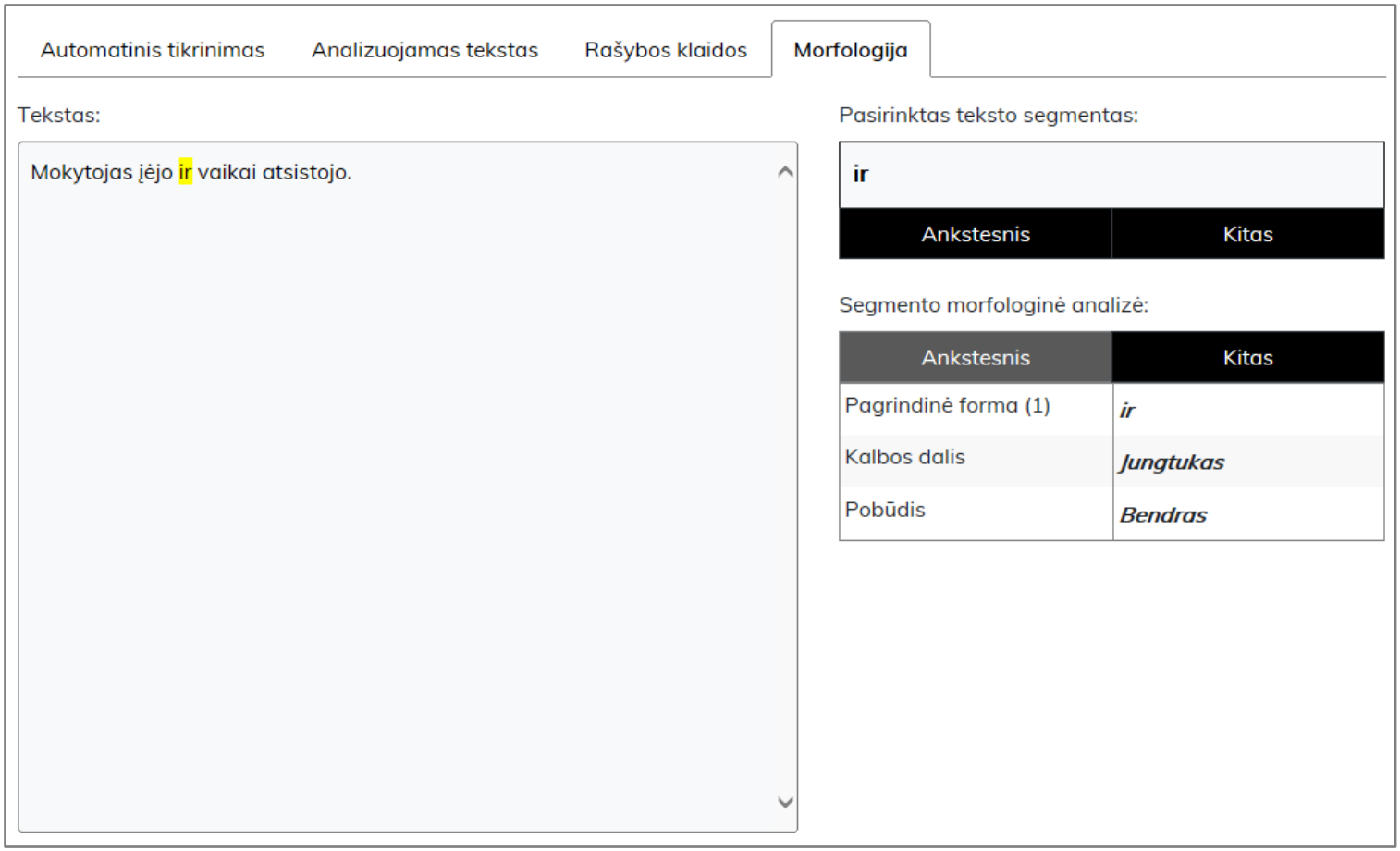
12	pretenduojančio		pretenduoti	VERB	vksm.dlv.veik.es.vyr.vns.K.	Case=Gen|Gender=Masc|Number=Sing|Tense=Pres|VerbForm=Part|Voice=Act	18	Atr	18:Atr	_			
13	tapti	tapti	VERB	vksm.bndr.	VerbForm=Inf	12	Obj	12:Obj	_					
14	didžiausiu	didis	ADJ	bdv.aukšč.vyr.vns.Įn.	Case=Ins|Degree=Sup|Gender=Masc|Number=Sing	13	Obj	13:Obj	_					
15	Baltijos	Baltija	NOUN	dkt.tikr.mot.vns.K.	Case=Gen|Gender=Fem|Number=Sing	16	Atr	16:Atr	_					
16	šalyse	šalis	NOUN	dkt.mot.dgs.Vt.	Case=Loc|Gender=Fem|Number=Plur	14	Adv	14:Adv	_					
17	logistikos	logistika	NOUN	dkt.mot.vns.K.	Case=Gen|Gender=Fem|Number=Sing	18	Atr	18:Atr	_					
18	komplekso	kompleksas	NOUN	dkt.vyr.vns.K.	Case=Gen|Gender=Masc|Number=Sing	19	Atr	19:Atr	_					
19	statinys	statinys	NOUN	dkt.vyr.vns.V.	Case=Nom|Gender=Masc|Number=Sing	10	Sub	10:Sub	SpaceAfter=No					
20	.	.	PUNCT	skyr.	_	10	AuxK	10:A						

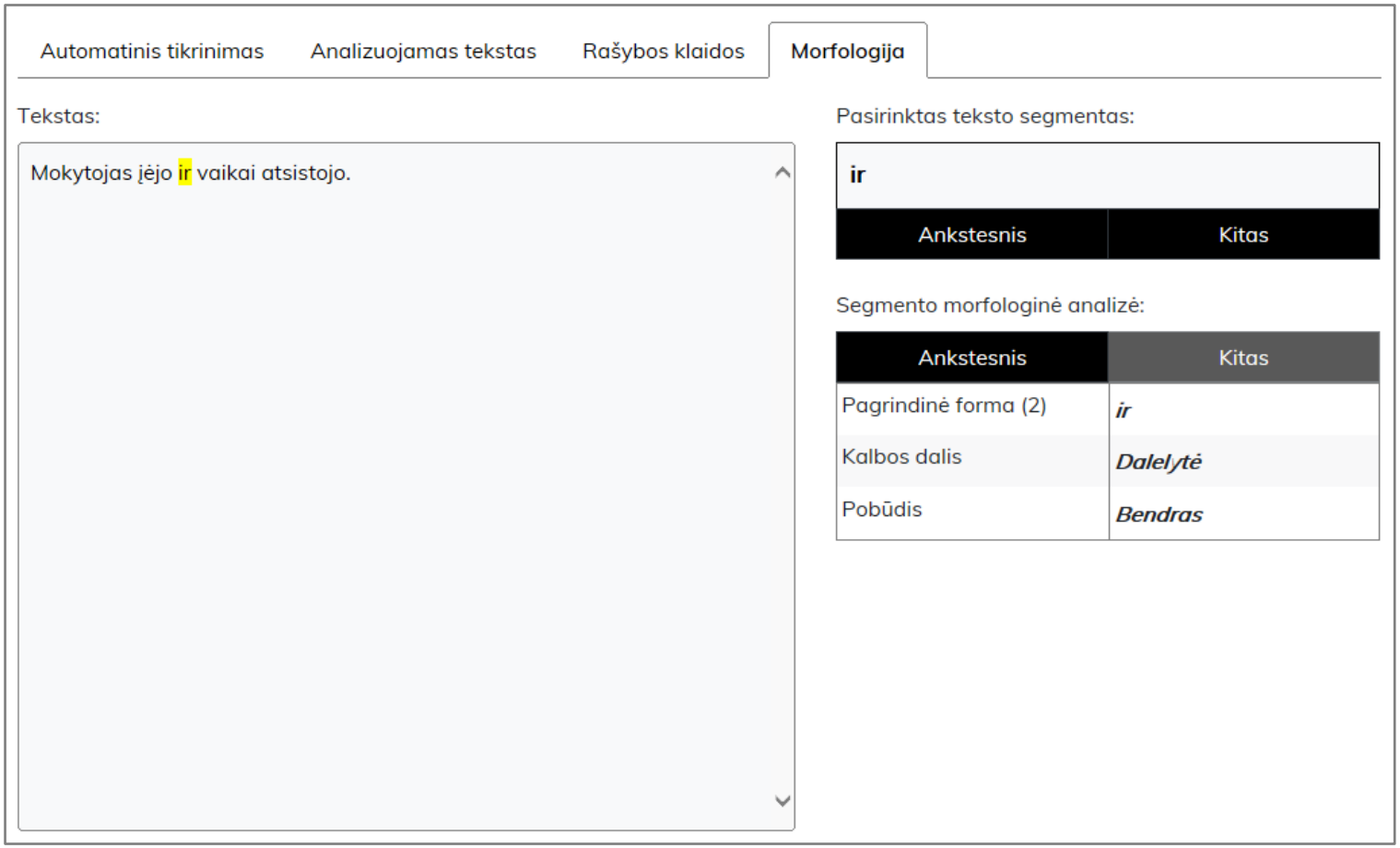
6 PRIEDAS: Sakinio „Mokytojas įėjo ir vaikai atsistojo“ morfologinė analizė, atlikta su *semantika.lt*

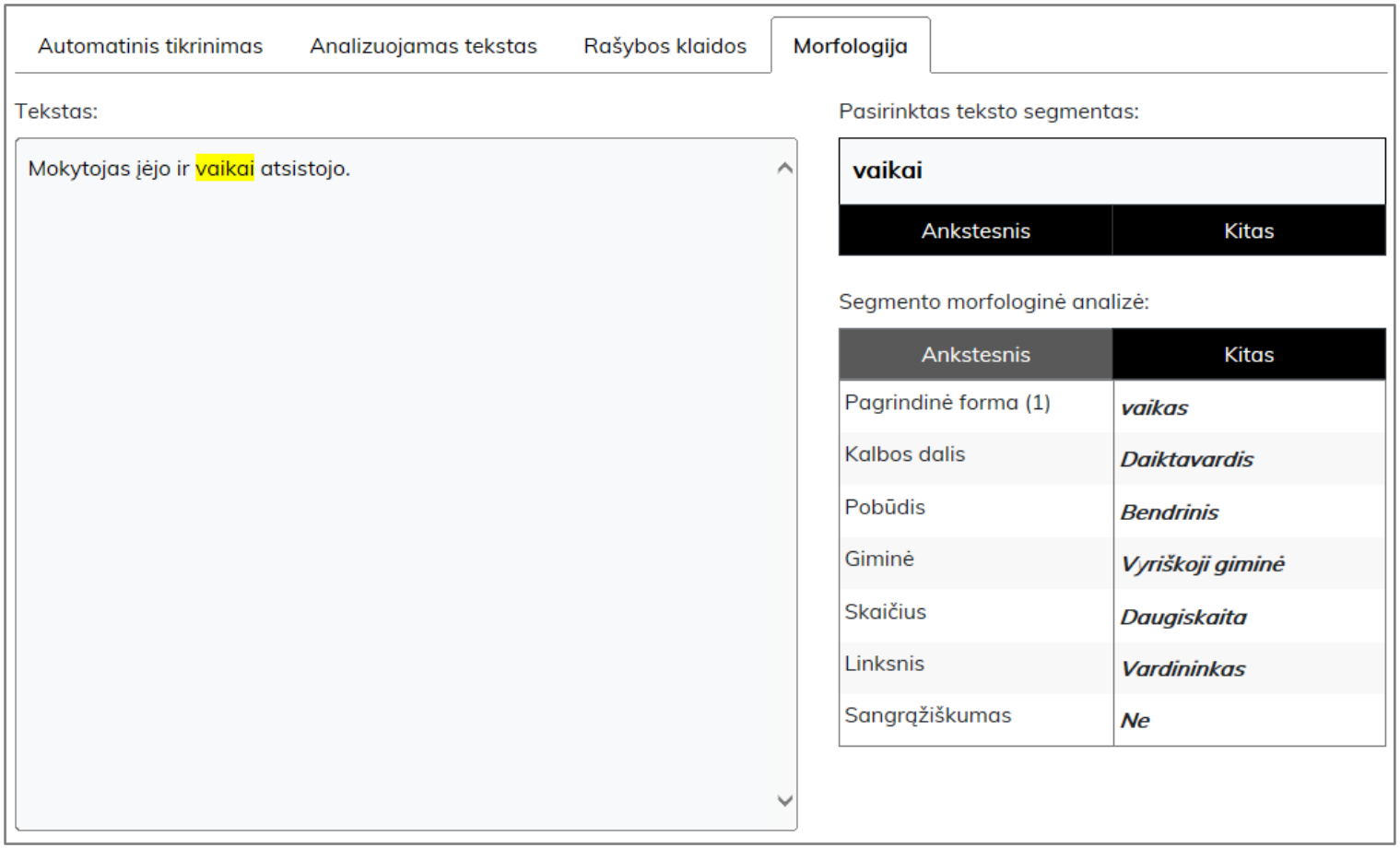


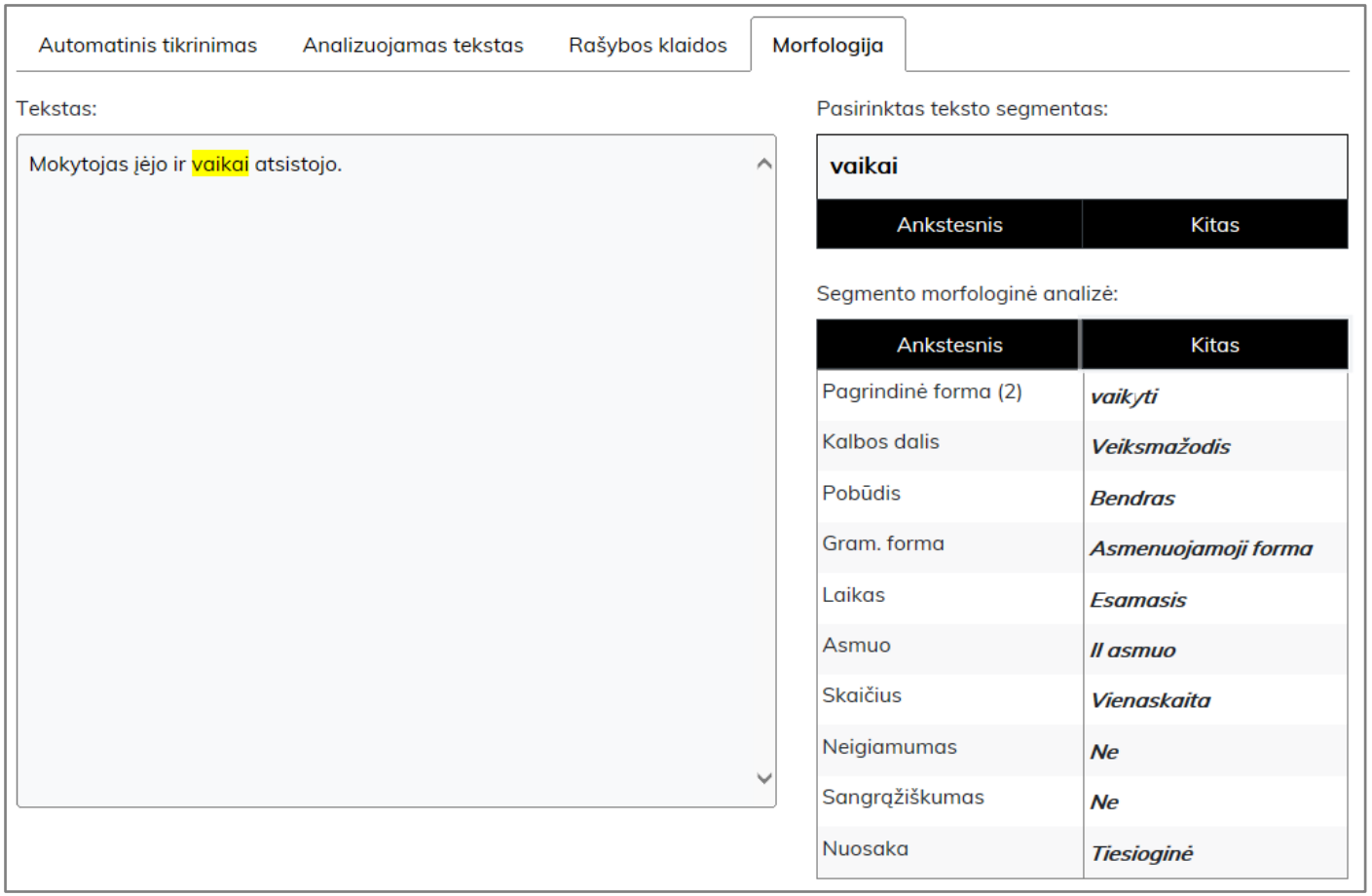

Automatinis tikrinimas
Analizuojamas tekstas
Rašybos klaidos
Morfologija
Tekstas:
Mokytojas įėjo ir vaikai atsistojo.
Pasirinktas teksto segmentas:
Mokytojas
Ankstesnis
Kitas
Segmento morfologinė analizė:
Ankstesnis
Kitas
Pagrindinė forma (2)
mokytoja
Kalbos dalis
Daiktavardis
Pobūdis
Bendrinis
Giminė
Moteriškoji giminė
Skaičius
Daugiskaita
Linksnis
Galininkas
Sangrąžiškumas
Ne

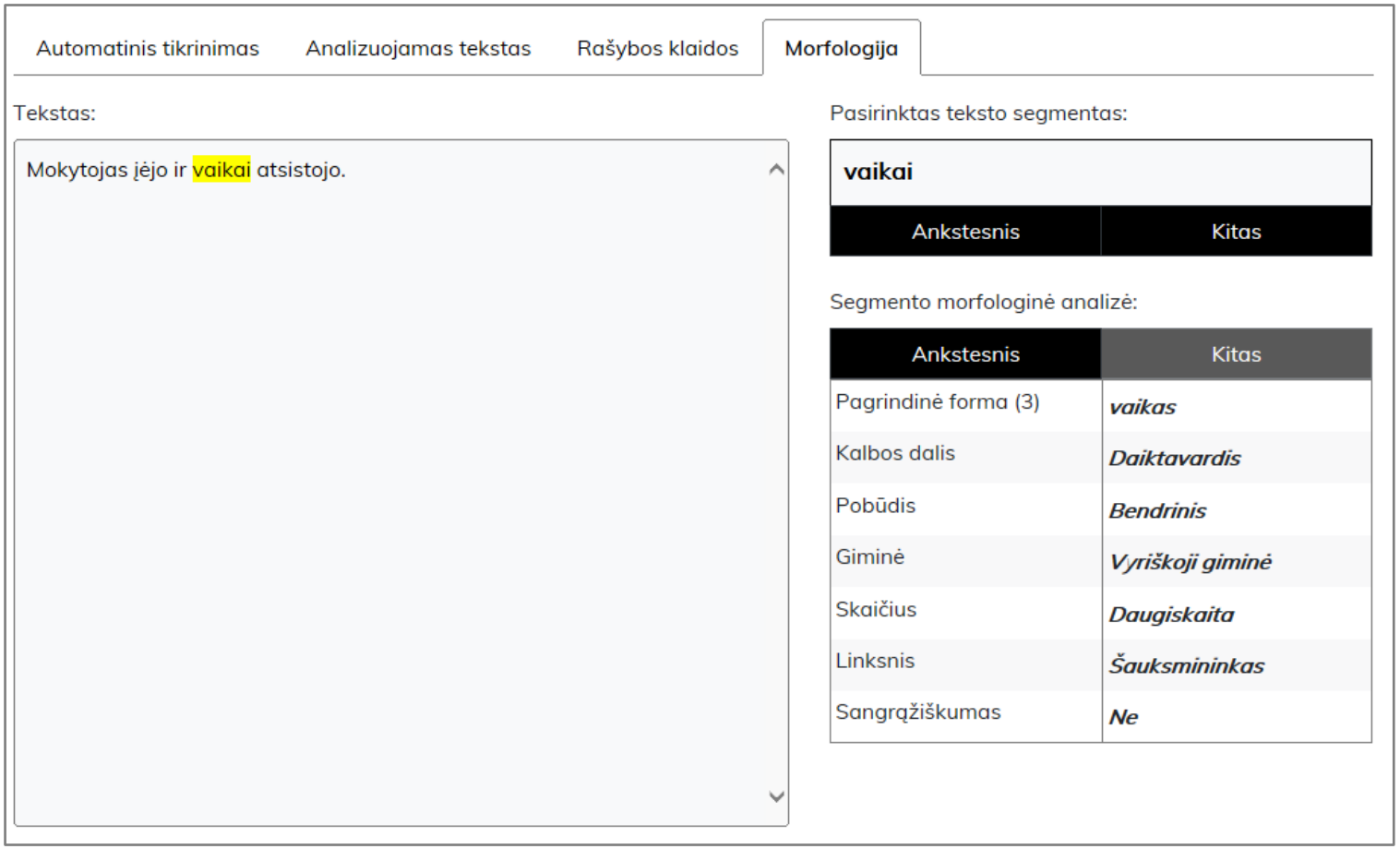

Automatinis tikrinimas
Analizuojamas tekstas
Rašybos klaidos
Morfologija
Tekstas:
Mokytojas įėjo ir vaikai atsistojo.
Pasirinktas teksto segmentas:
įėjo
Ankstesnis
Kitas
Segmento morfologinė analizė:
Ankstesnis
Kitas
Pagrindinė forma (1)
įeiti
Kalbos dalis
Veiksmažodis
Pobūdis
Bendras
Gram. forma
Asmenuojamoji forma
Laikas
Būtasis
Asmuo
III asmuo
Neigiamumas
Ne
Sangrąžiškumas
Ne
Nuosaka
Tiesioginė

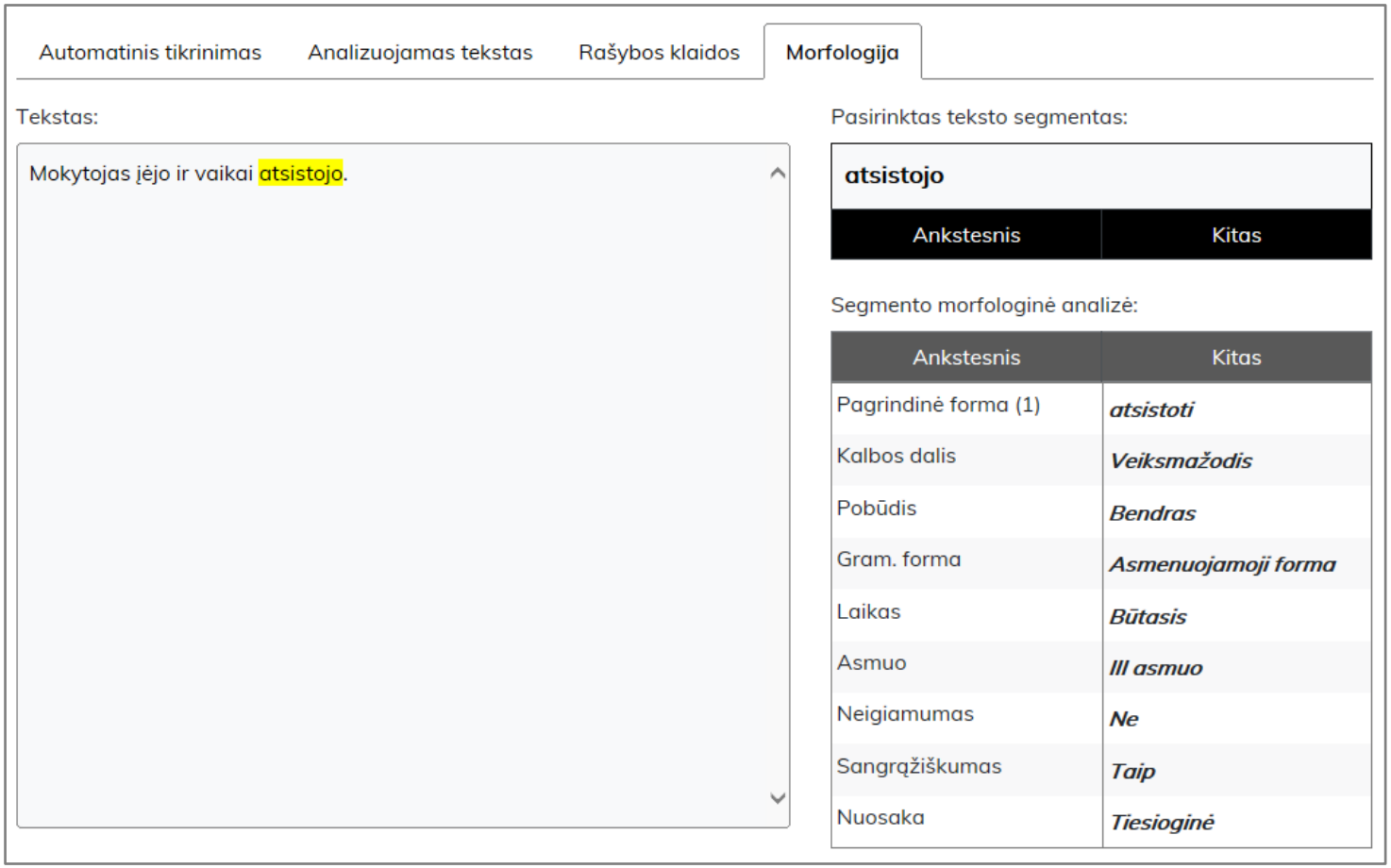

Automatinis tikrinimas
Analizuojamas tekstas
Rašybos klaidos
Morfologija
Tekstas:
Mokytojas įėjo ir vaikai atsistojo.
Pasirinktas teksto segmentas:
ir
Ankstesnis
Kitas
Segmento morfologinė analizė:
Ankstesnis
Kitas
Pagrindinė forma (1)
ir
Kalbos dalis
Jungtukas
Pobūdis
Bendras


Automatinis tikrinimas
Analizuojamas tekstas
Rašybos klaidos
Morfologija
Tekstas:
Mokytojas įėjo ir vaikai atsistojo.
Pasirinktas teksto segmentas:
ir
Ankstesnis
Kitas
Segmento morfologinė analizė:
Ankstesnis
Kitas
Pagrindinė forma (2)
ir
Kalbos dalis
Dalelytė
Pobūdis
Bendras

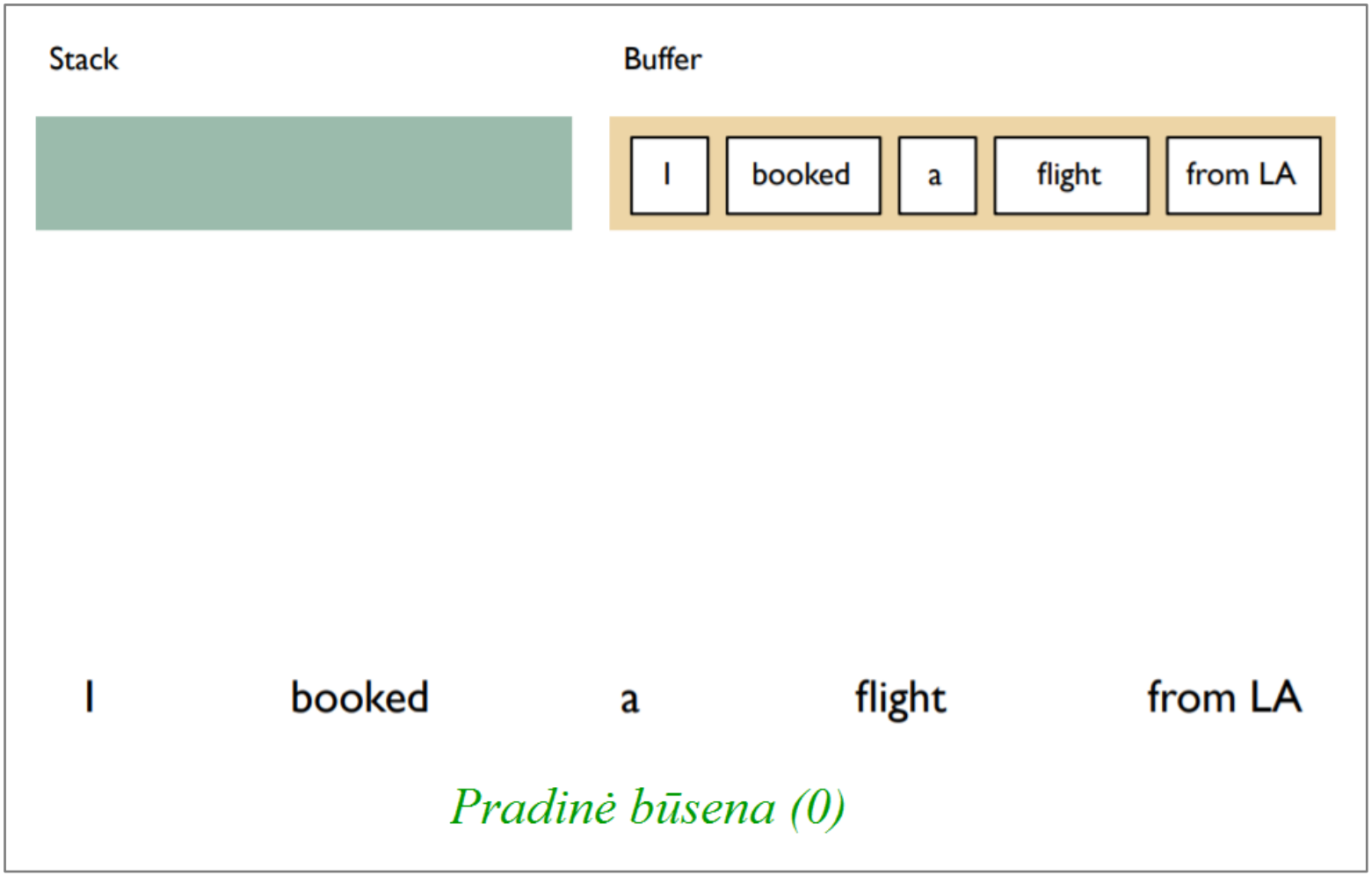

Automatinis tikrinimas
Analizuojamas tekstas
Rašybos klaidos
Morfologija
Tekstas:
Mokytojas įėjo ir vaikai atsistojo.
Pasirinktas teksto segmentas:
vaikai
Ankstesnis
Kitas
Segmento morfologinė analizė:
Ankstesnis
Kitas
Pagrindinė forma (1)
vaikas
Kalbos dalis
Daiktavardis
Pobūdis
Bendrinis
Giminė
Vyriškoji giminė
Skaičius
Daugiskaita
Linksnis
Vardininkas
Sangrąžiškumas
Ne

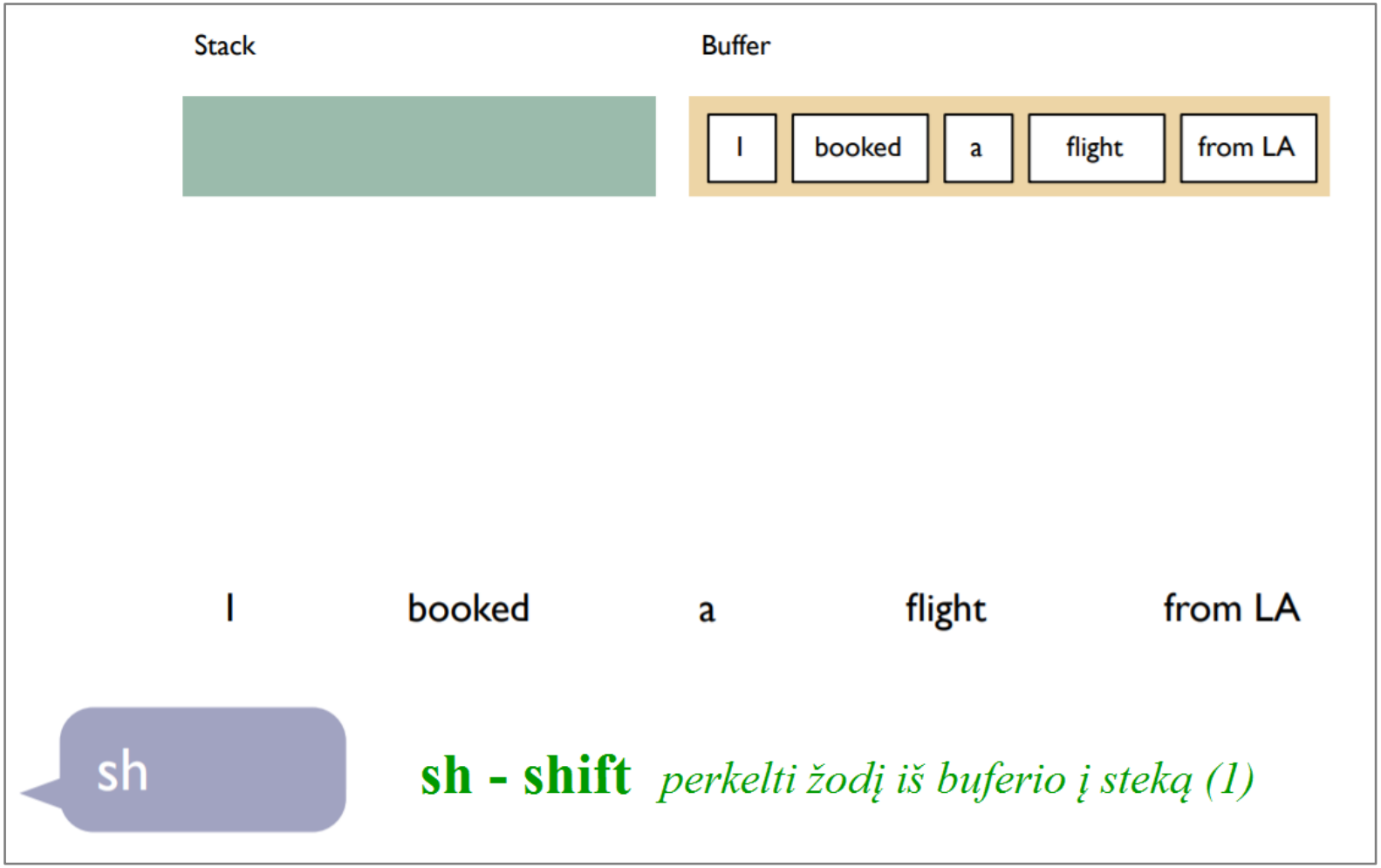

Automatinis tikrinimas
Analizuojamas tekstas
Rašybos klaidos
Morfologija
Tekstas:
Mokytojas įėjo ir vaikai atsistojo.
Pasirinktas teksto segmentas:
vaikai
Ankstesnis
Kitas
Segmento morfologinė analizė:
Ankstesnis
Kitas
Pagrindinė forma (2)
vaikyti
Kalbos dalis
Veiksmažodis
Pobūdis
Bendras
Gram. forma
Asmenuojamoji forma
Laikas
Esamasis
Asmuo
II asmuo
Skaičius
Vienaskaita
Neigiamumas
Ne
Sangrąžiškumas
Ne
Nuosaka
Tiesioginė


Automatinis tikrinimas
Analizuojamas tekstas
Rašybos klaidos
Morfologija
Tekstas:
Mokytojas įėjo ir vaikai atsistojo.
Pasirinktas teksto segmentas:
vaikai
Ankstesnis
Kitas
Segmento morfologinė analizė:
Ankstesnis
Kitas
Pagrindinė forma (3)
vaikas
Kalbos dalis
Daiktavardis
Pobūdis
Bendrinis
Giminė
Vyriškoji giminė
Skaičius
Daugiskaita
Linksnis
Šauksmininkas
Sangrąžiškumas
Ne

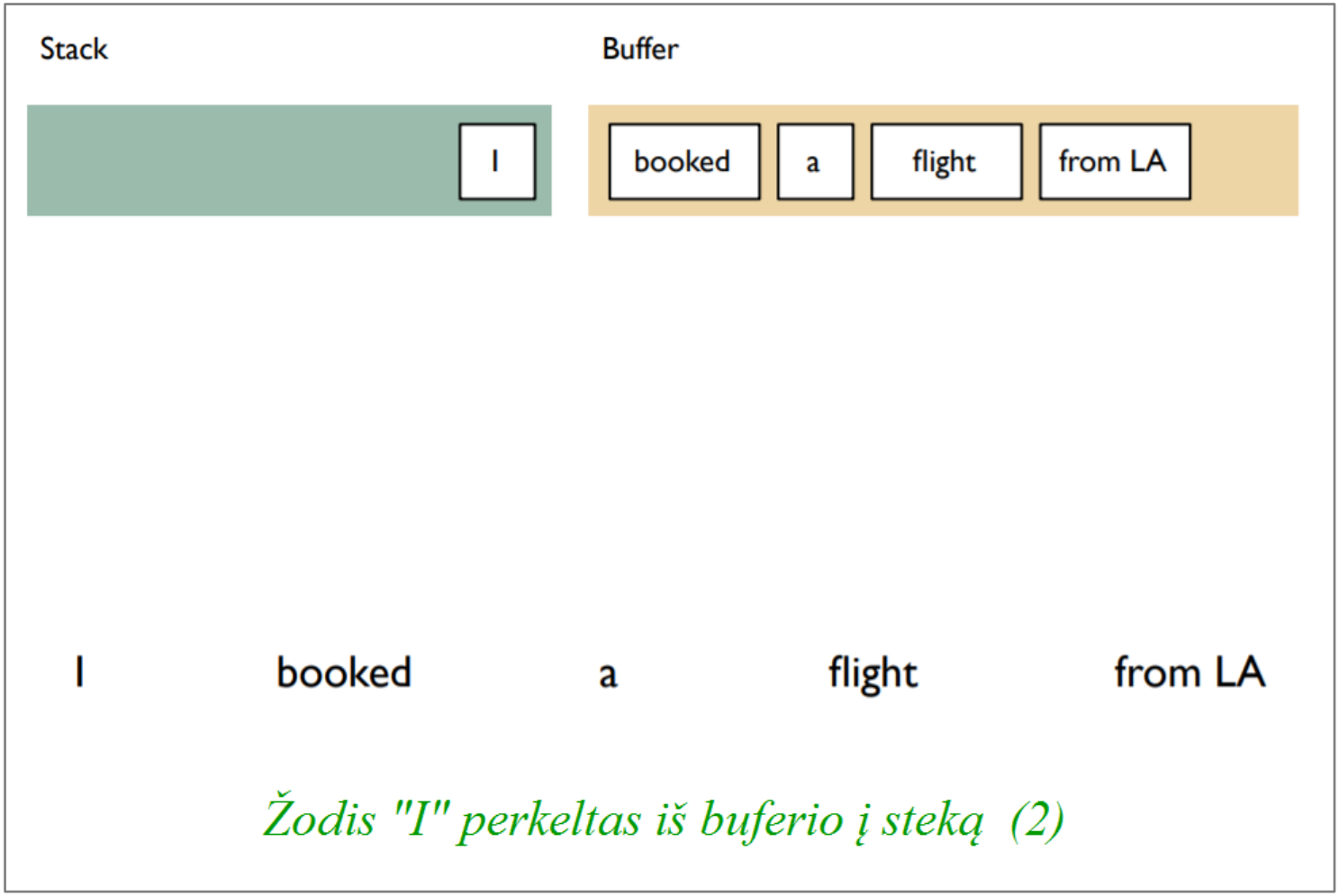

Automatinis tikrinimas
Analizuojamas tekstas
Rašybos klaidos
Morfologija
Tekstas:
Mokytojas įėjo ir vaikai atsistojo.
Pasirinktas teksto segmentas:
atsistojo
Ankstesnis
Kitas
Segmento morfologinė analizė:
Ankstesnis
Kitas
Pagrindinė forma (1)
atsistoti
Kalbos dalis
Veiksmažodis
Pobūdis
Bendras
Gram. forma
Asmenuojamoji forma
Laikas
Būtasis
Asmuo
III asmuo
Neigiamumas
Ne
Sangrąžiškumas
Taip
Nuosaka
Tiesioginė

7 PRIEDAS: Lankų algoritmo žingsniai

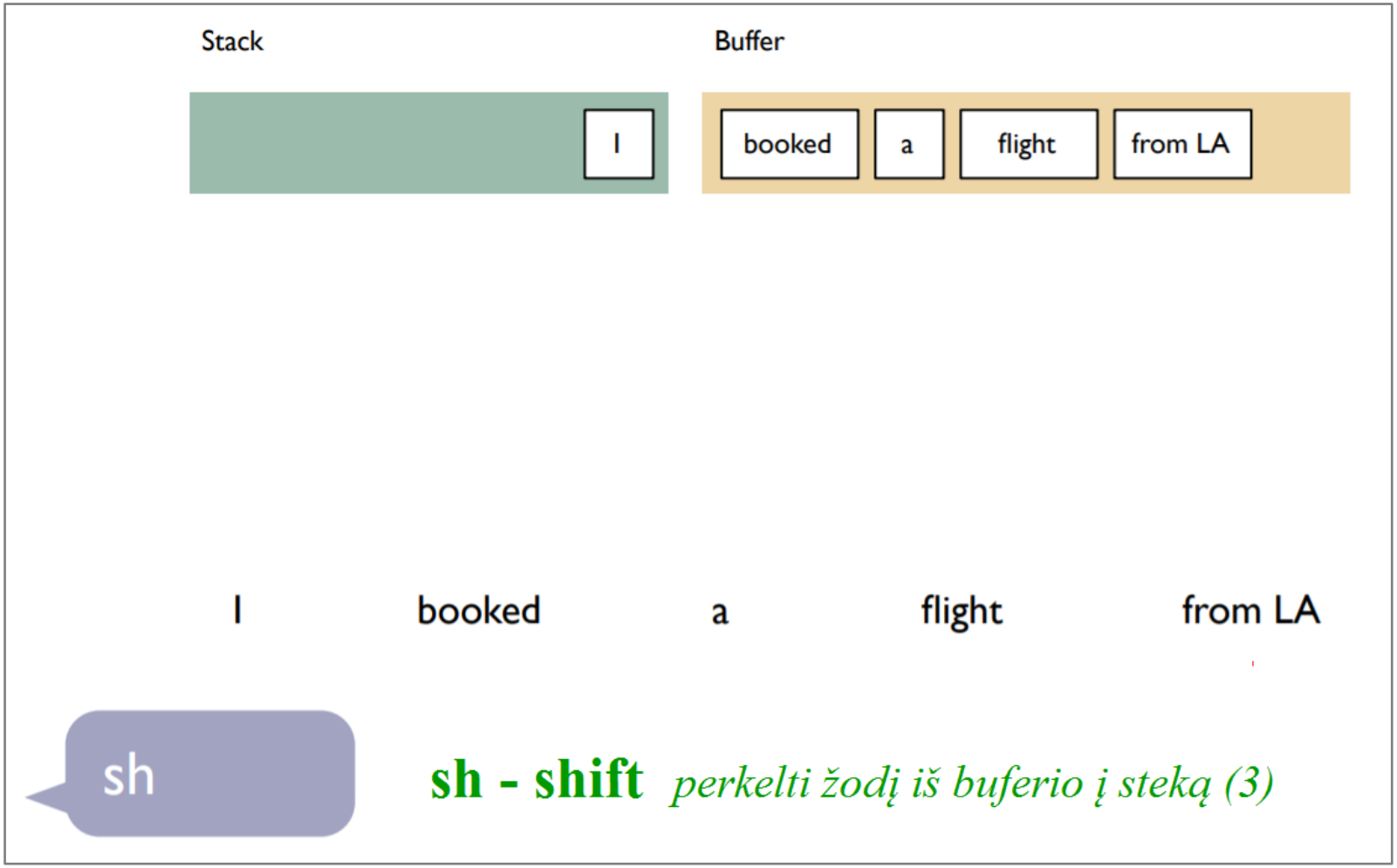


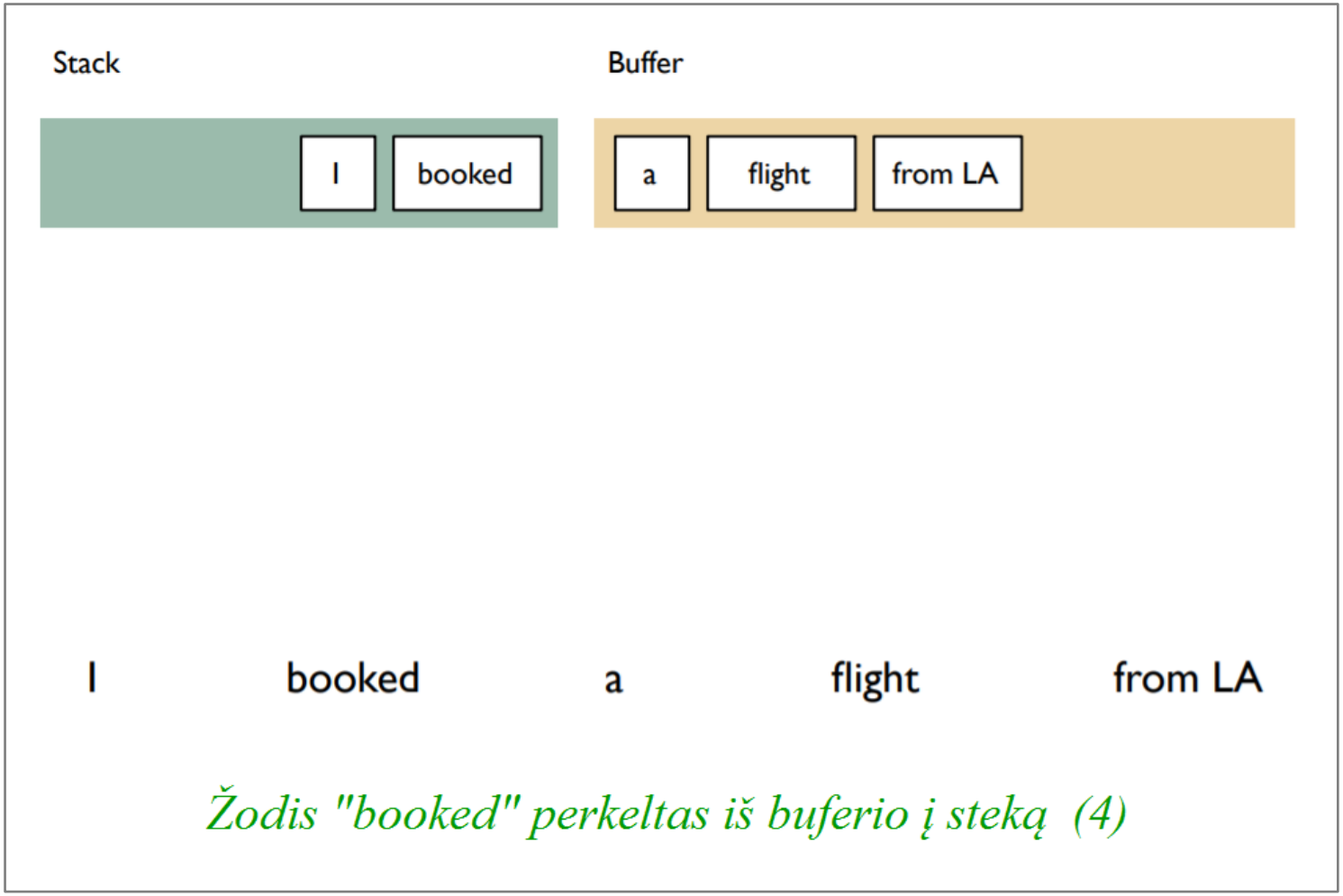


Žodžiai iš buferio į steką perkeliami tol, kol susidaro tokia žodžių grupė, iš kurios galima suformuoti žodžių junginį.

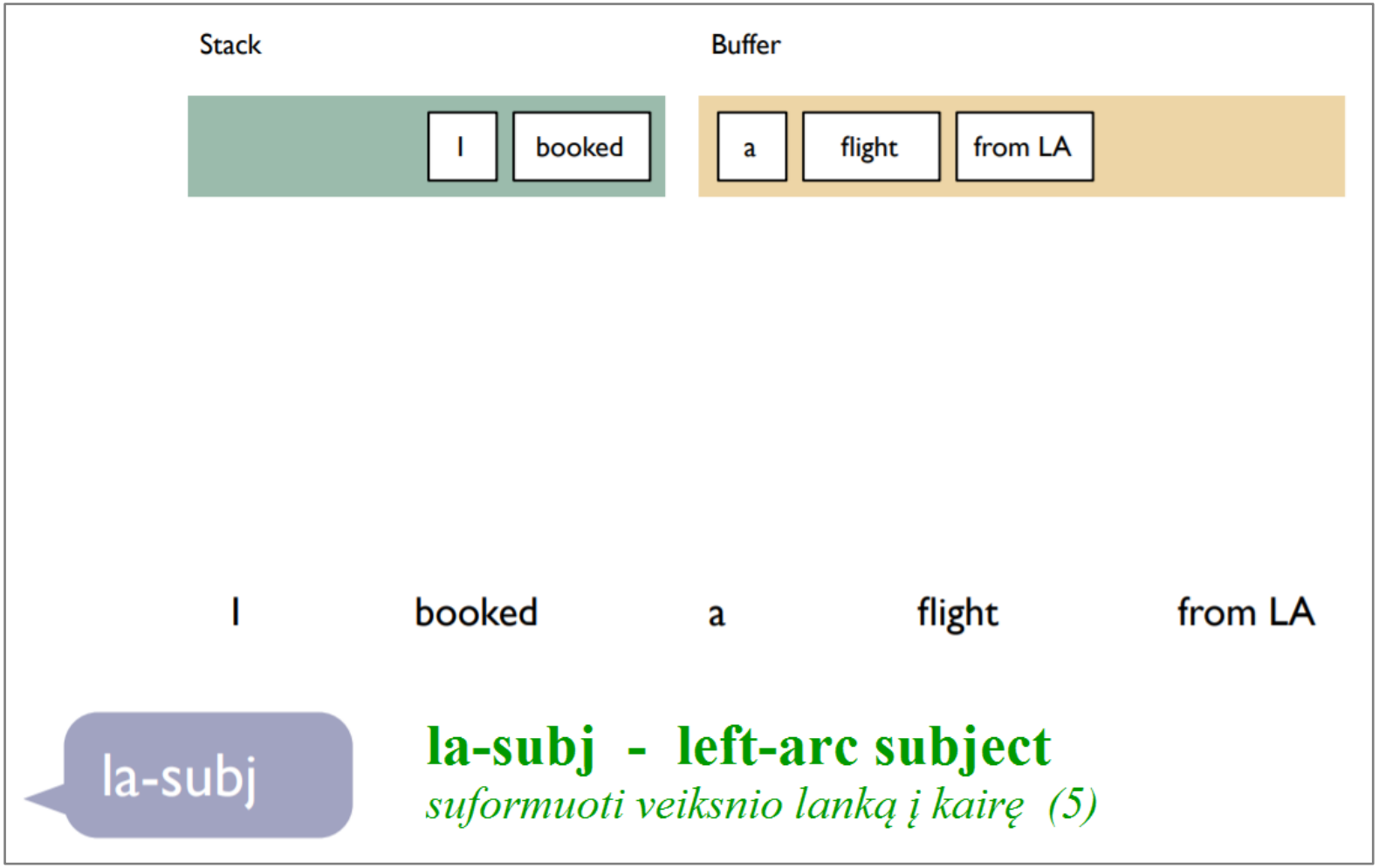


Kadangi iš steke esančio žodžio dar negalima suformuoti žodžių junginio, į jį perkeliamas dar vienas žodis.

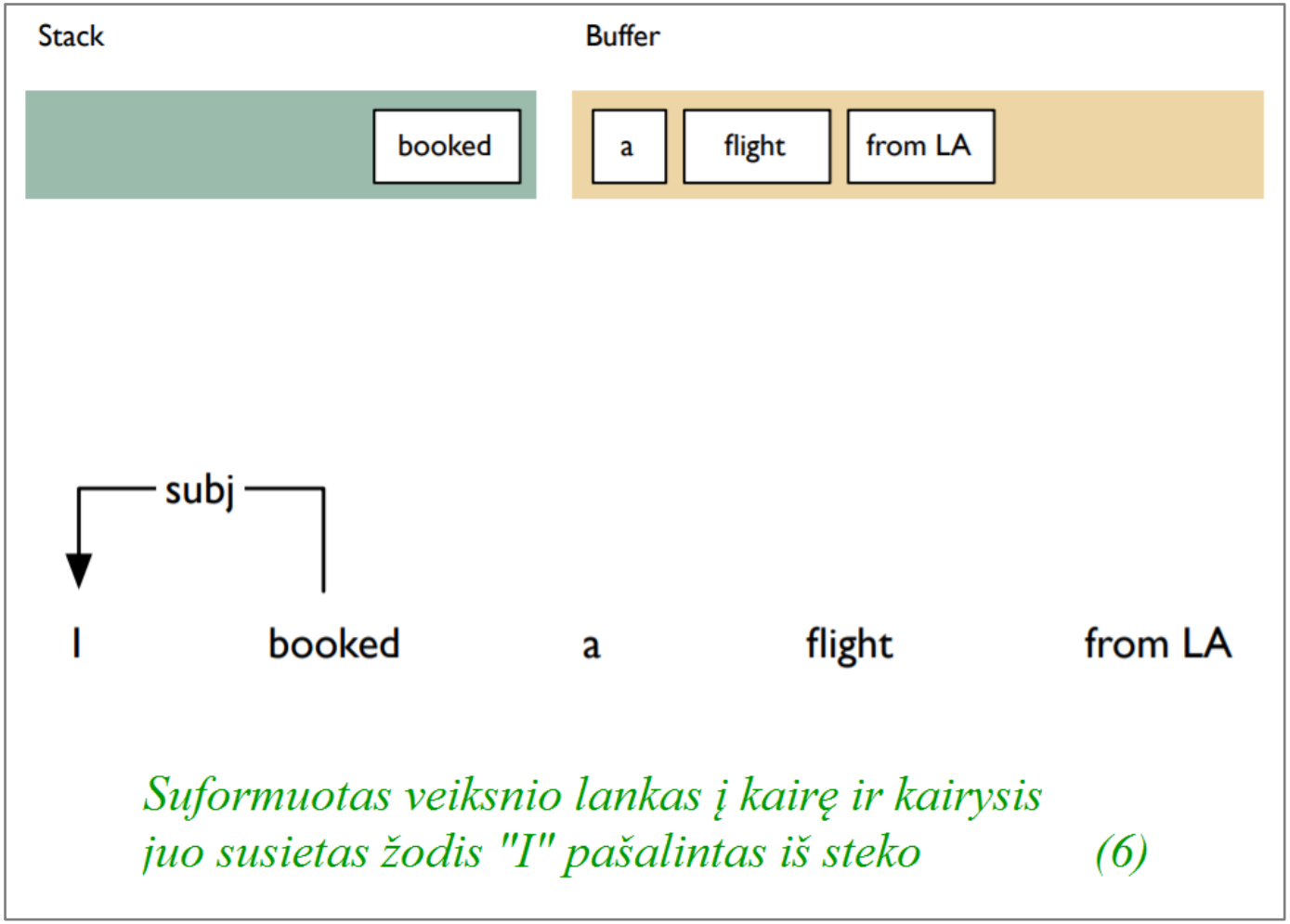

Stack
Buffer
I
booked
a
flight
from LA
I
booked
a
flight
from LA
sh
sh - shift perkelti žodį iš buferio į steką (3)



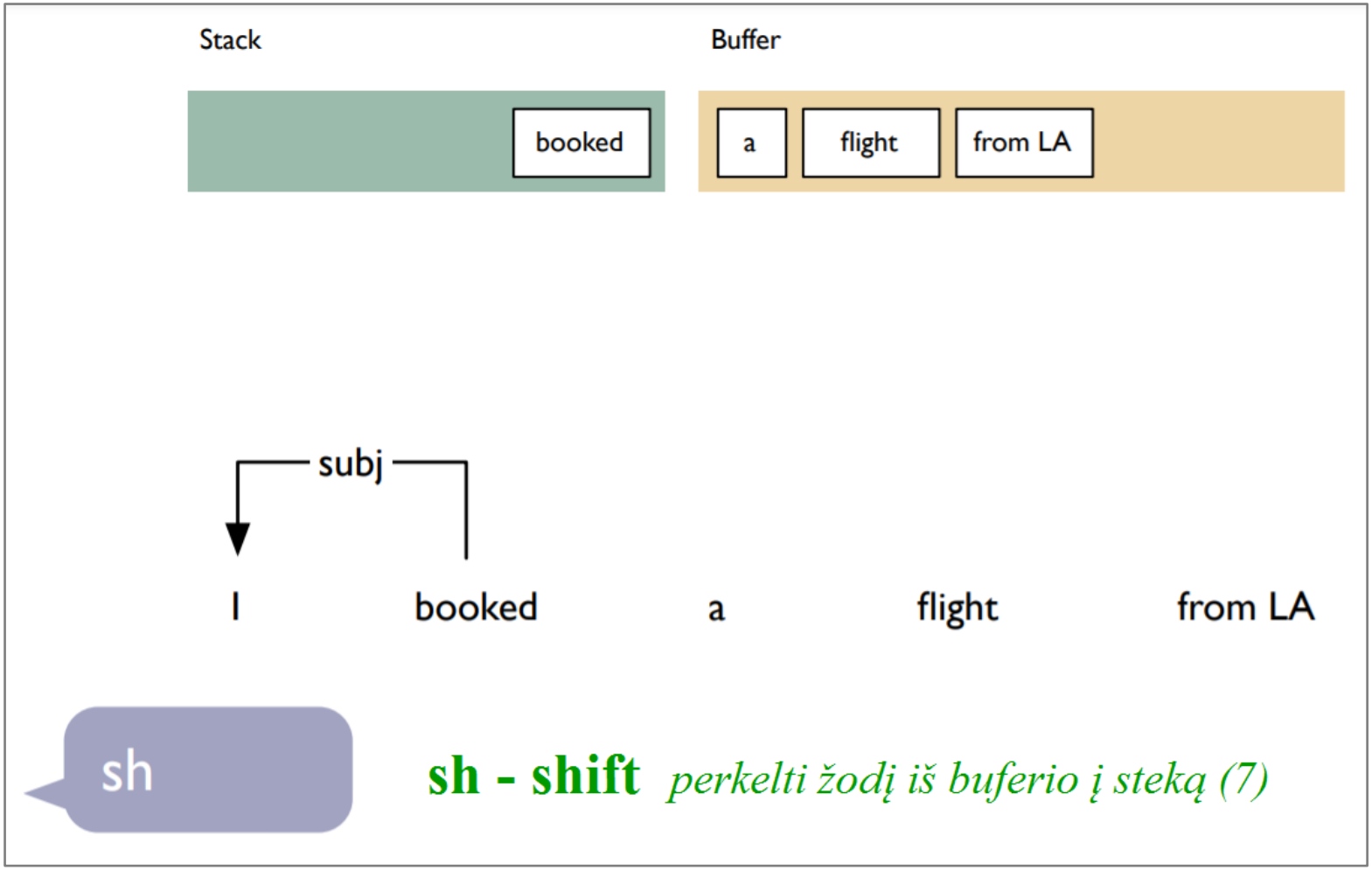
Žodis "booked" perkeltas iš buferio į steką (4)



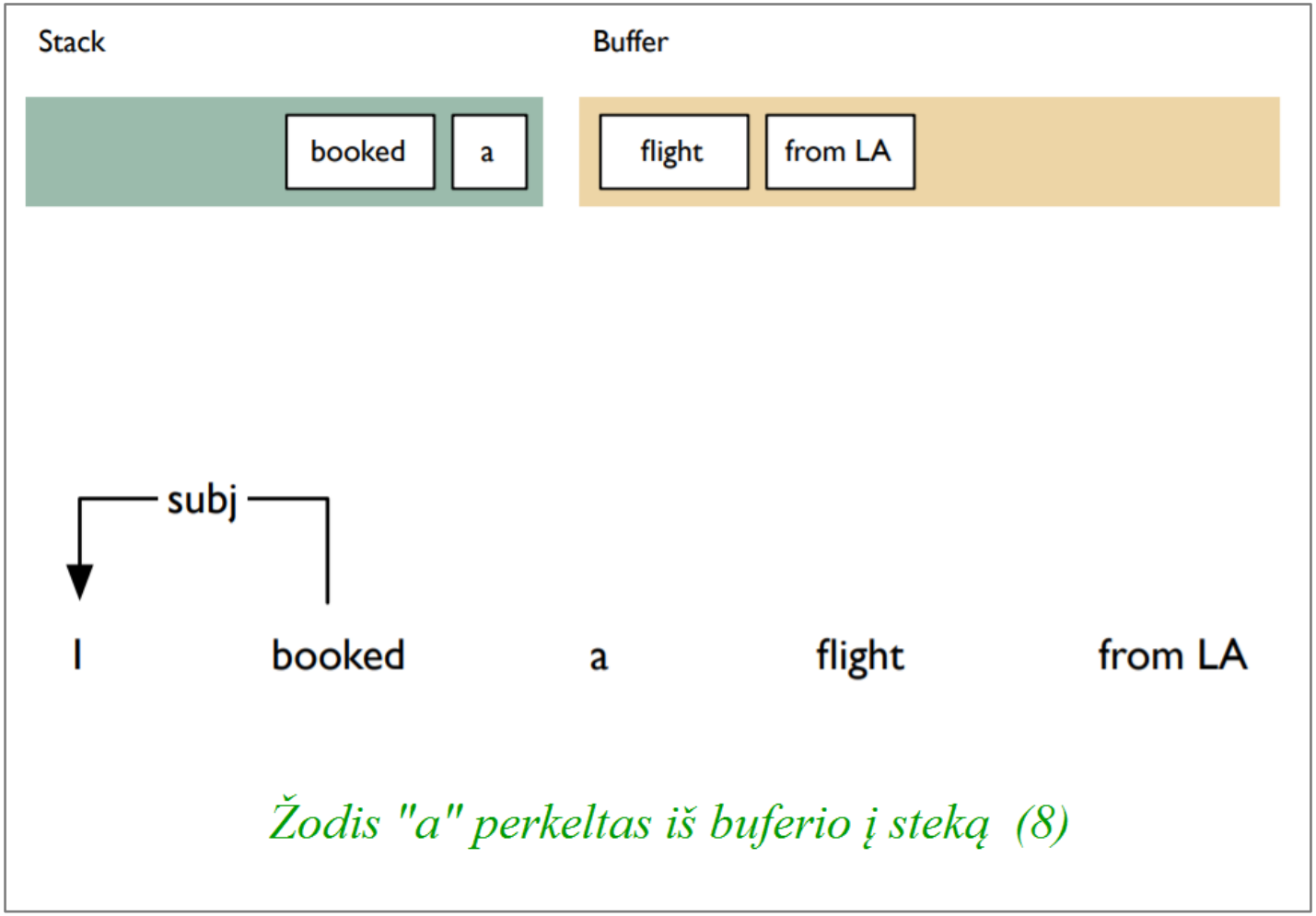
Kadangi steke esantys žodžiai jau gali sudaryti žodžių junginį, suformuojamas juos jungiantis lankas.



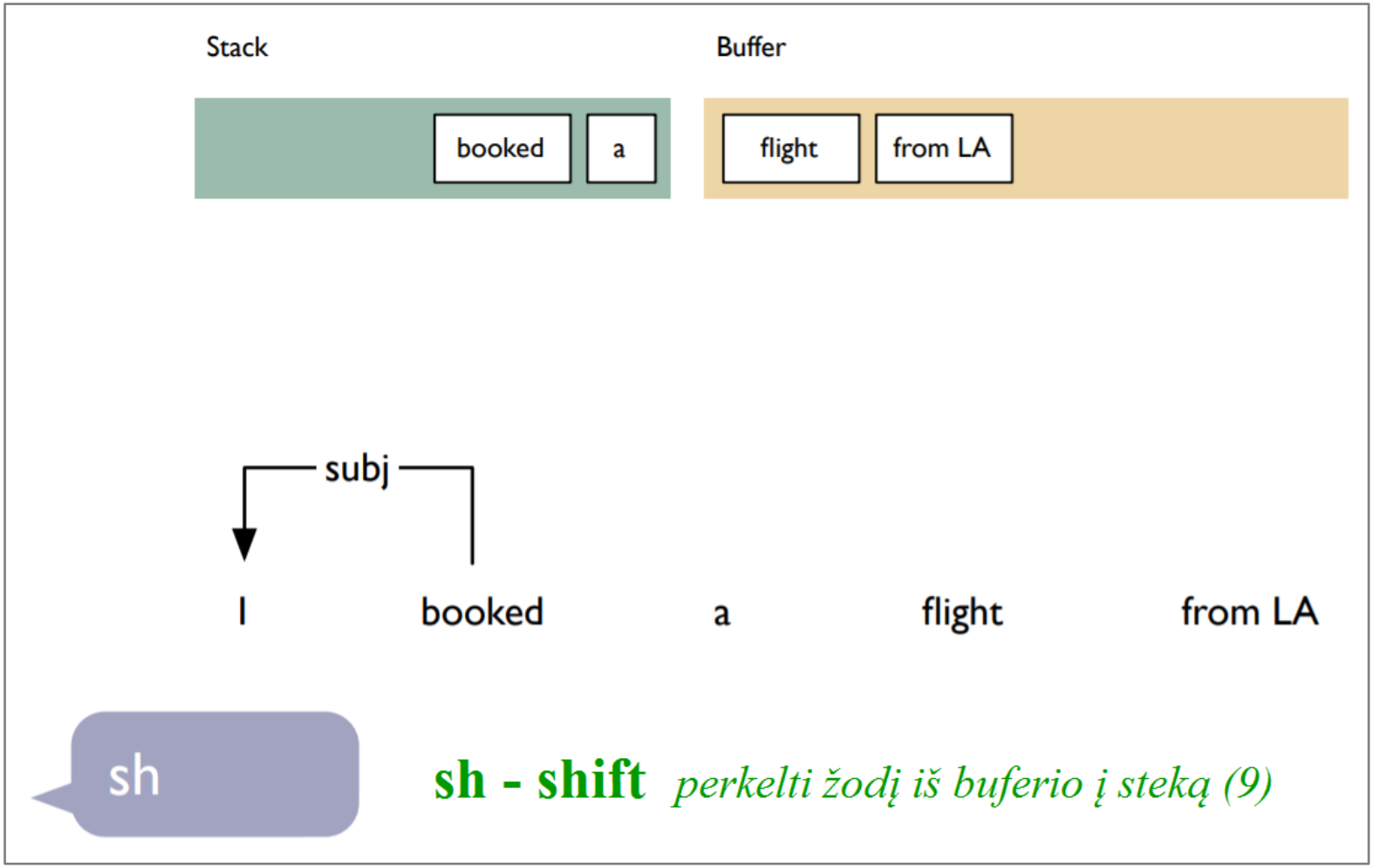
Suformavus lanką, kraštinis juo siejamas žodis (į kurį nukreipta lanko rodyklė) pašalinamas iš steko.

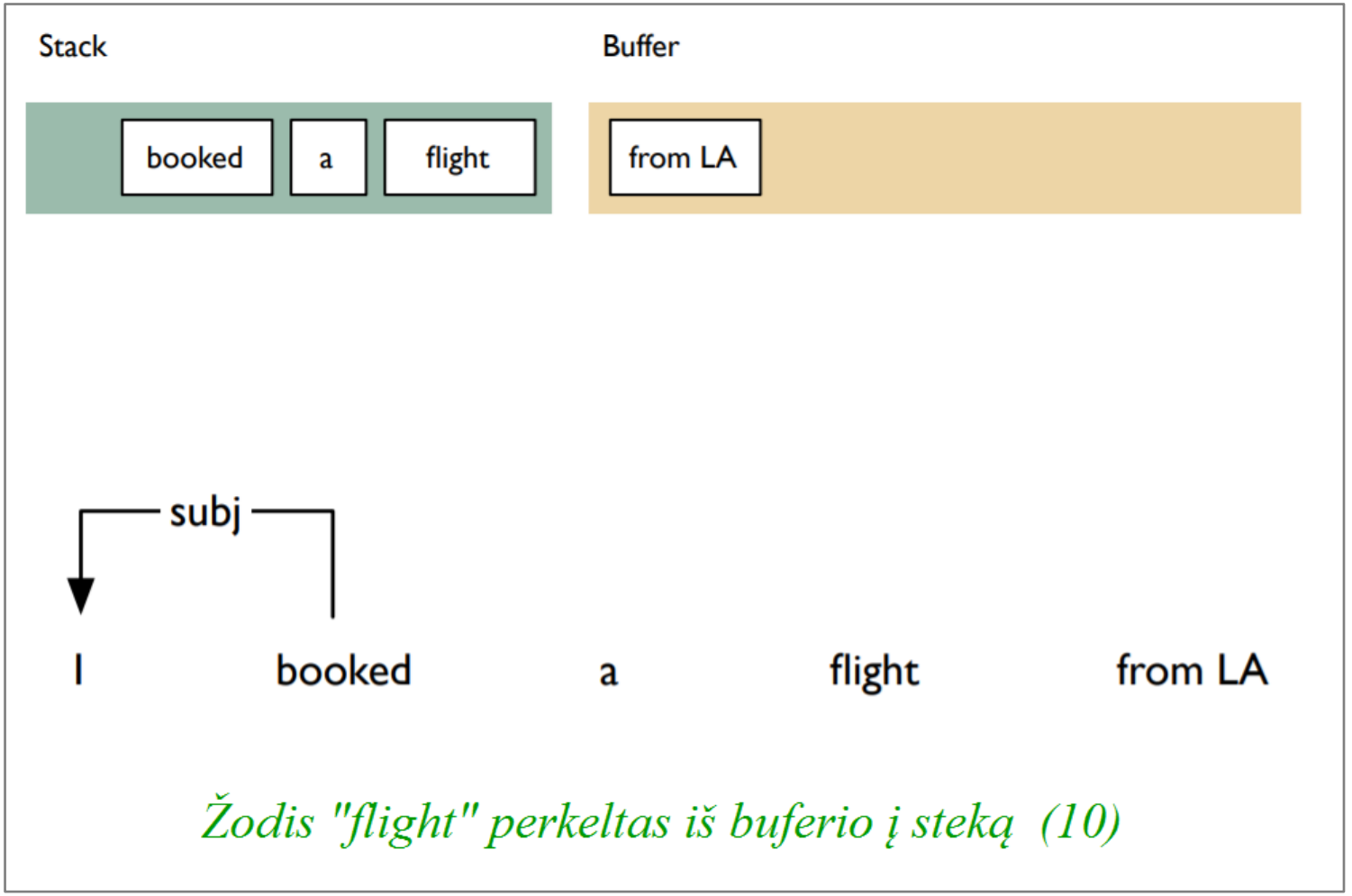


Steke liko vienas žodis, kuris negali sudaryti žodžių junginio, todėl į jį perkeliamas dar vienas žodis iš buferio.

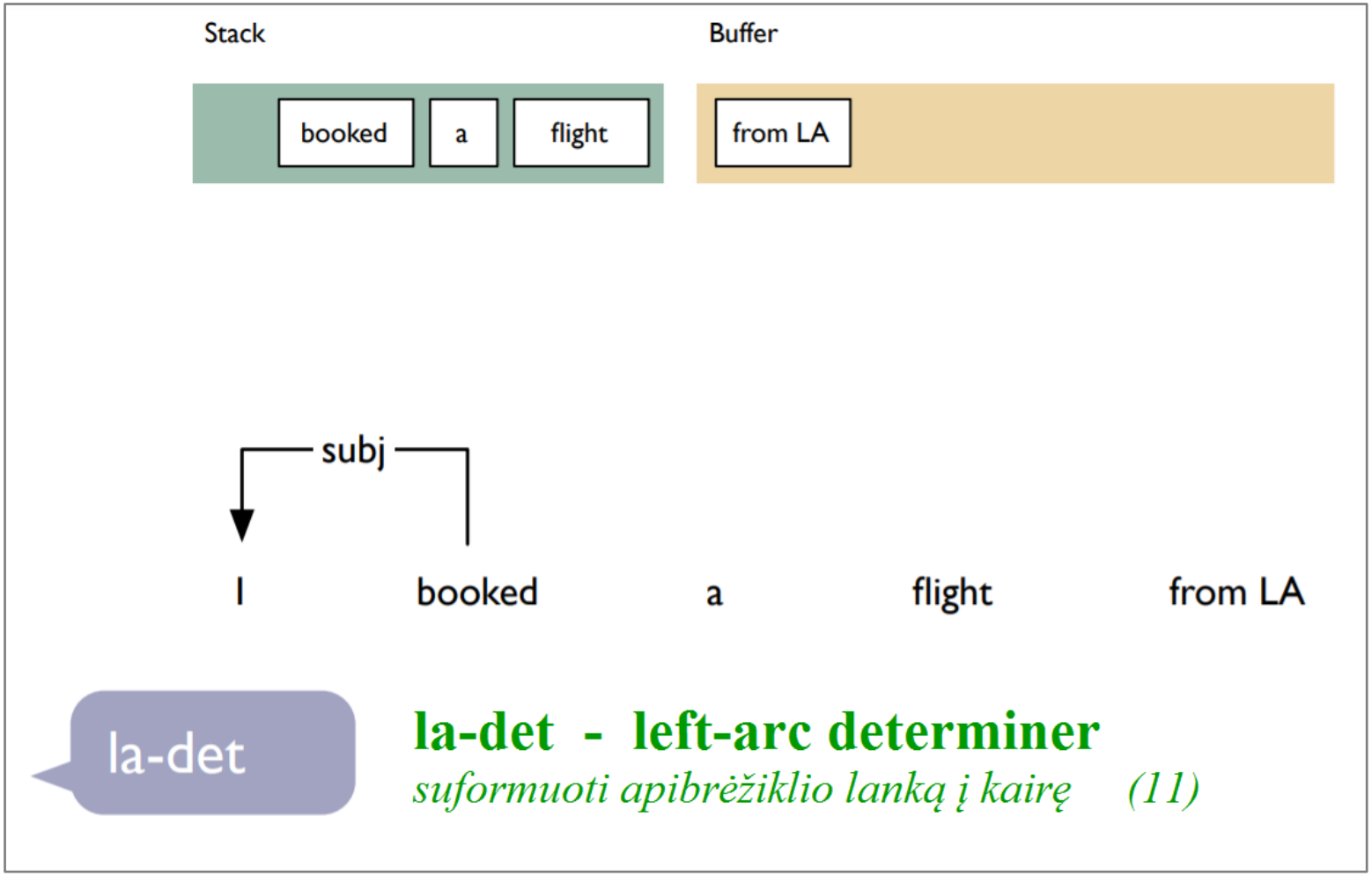


Steke esnatys du žodžiai nesudaro žodžių junginio, todėl į jį perkeliamas dar vienas žodis.

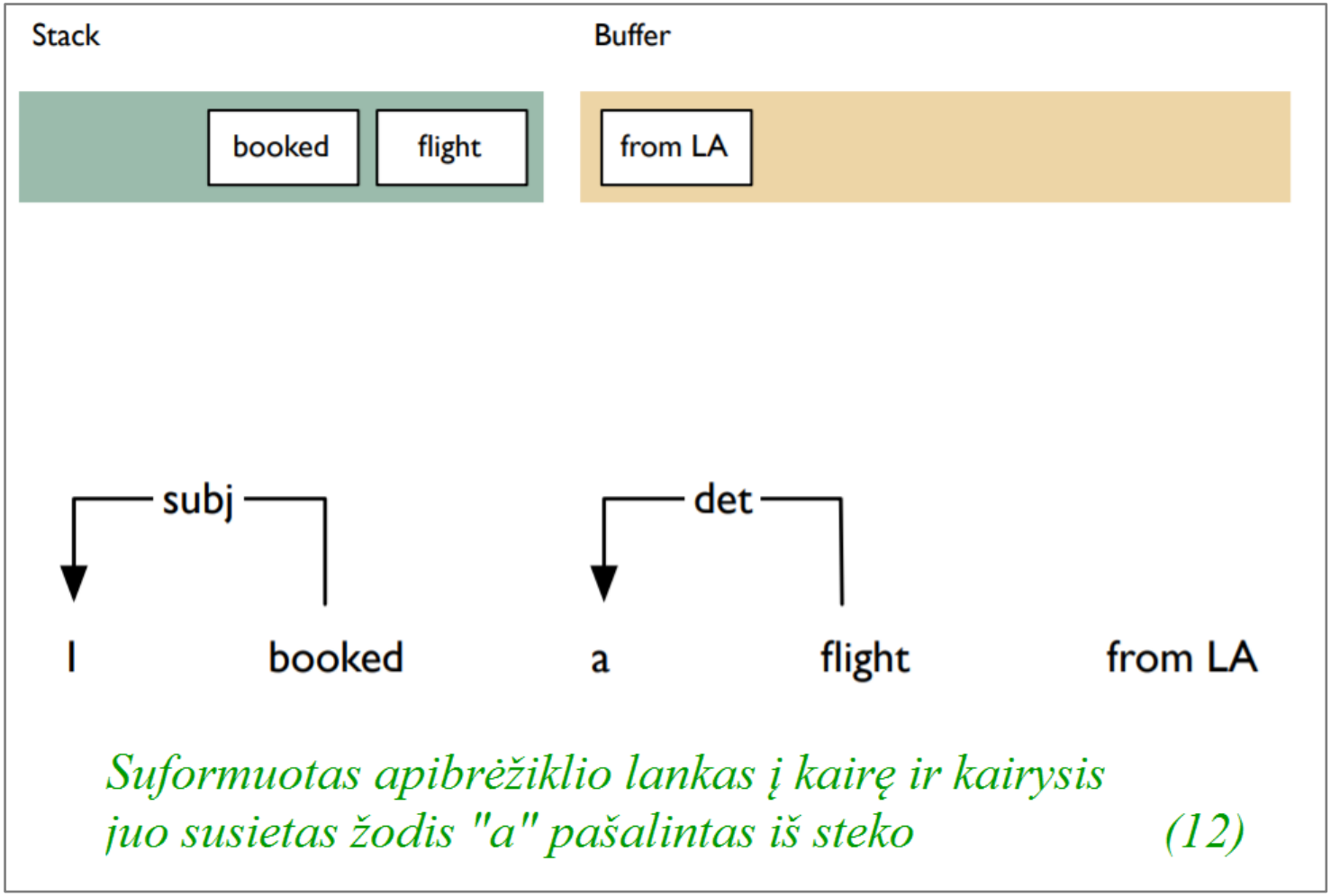

Stack
Buffer
booked
a
flight
from LA
subj
I
booked
a
flight
from LA
sh
sh - shift perkelti žodį iš buferio į steką (9)



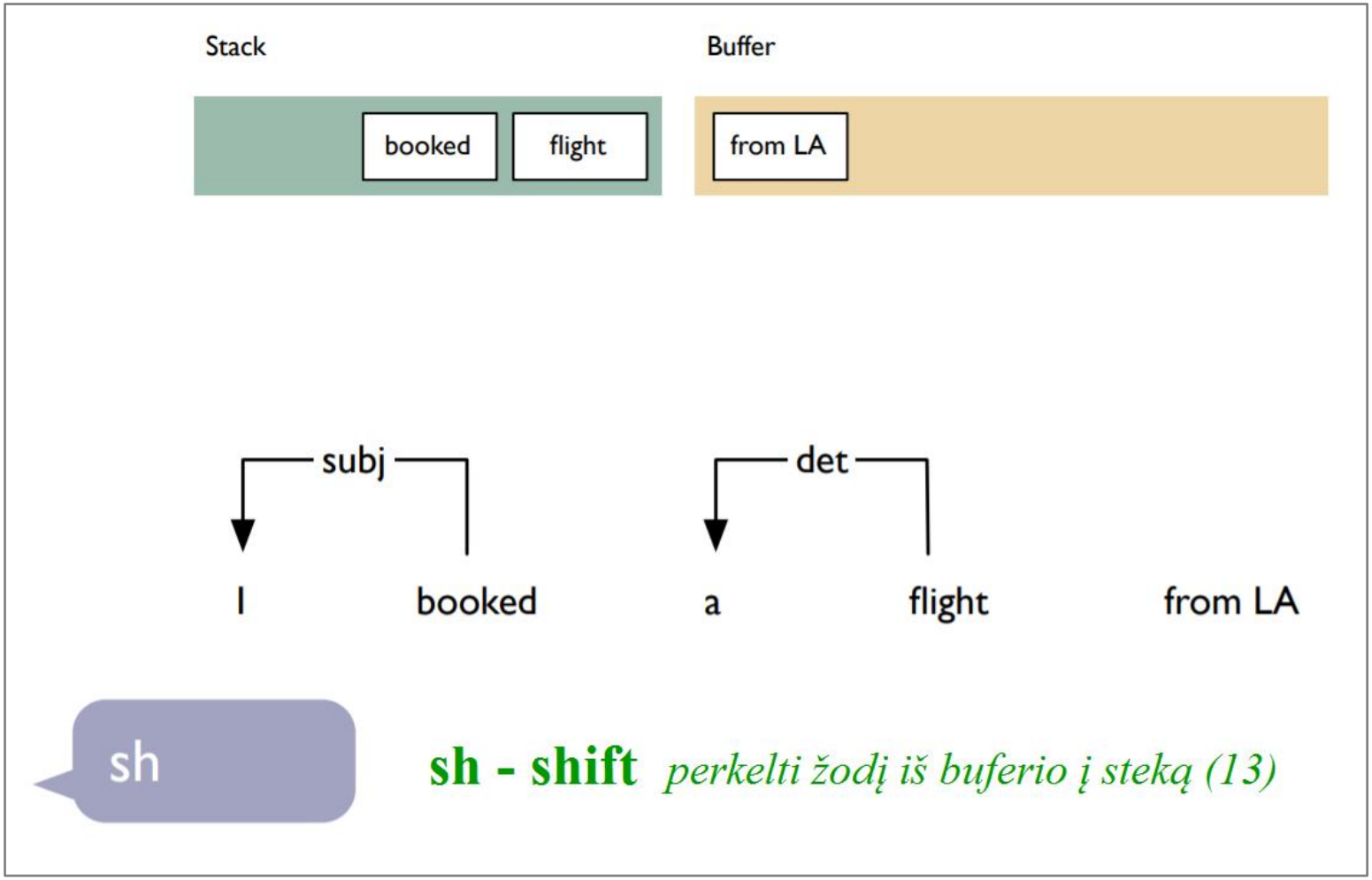
Žodis "flight" perkeltas iš buferio į steką (10)



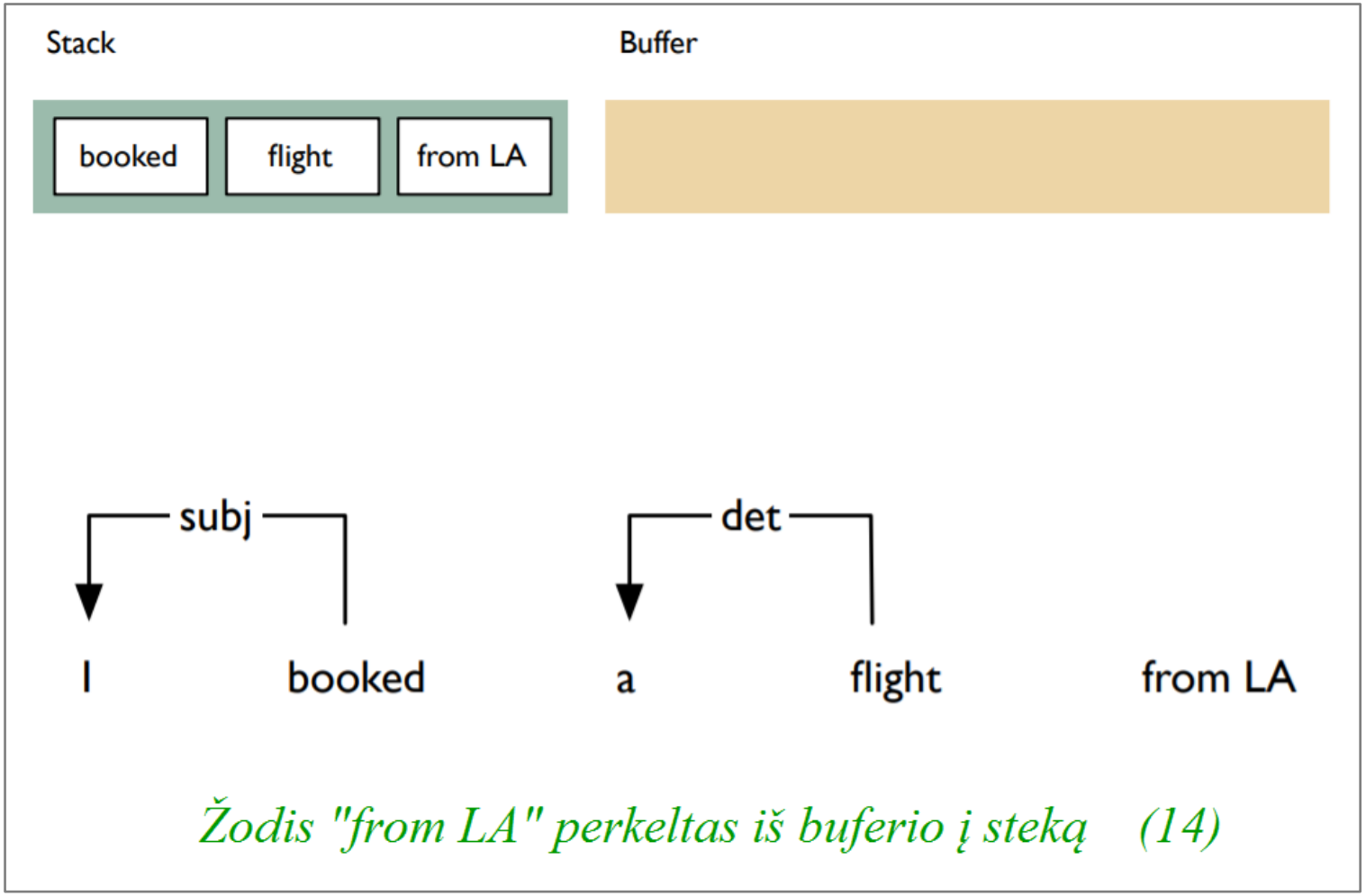
Du šalia esantys žodžiai gali sudaryti žodžių junginį, todėl formuojamas juos jungiantis lankas.



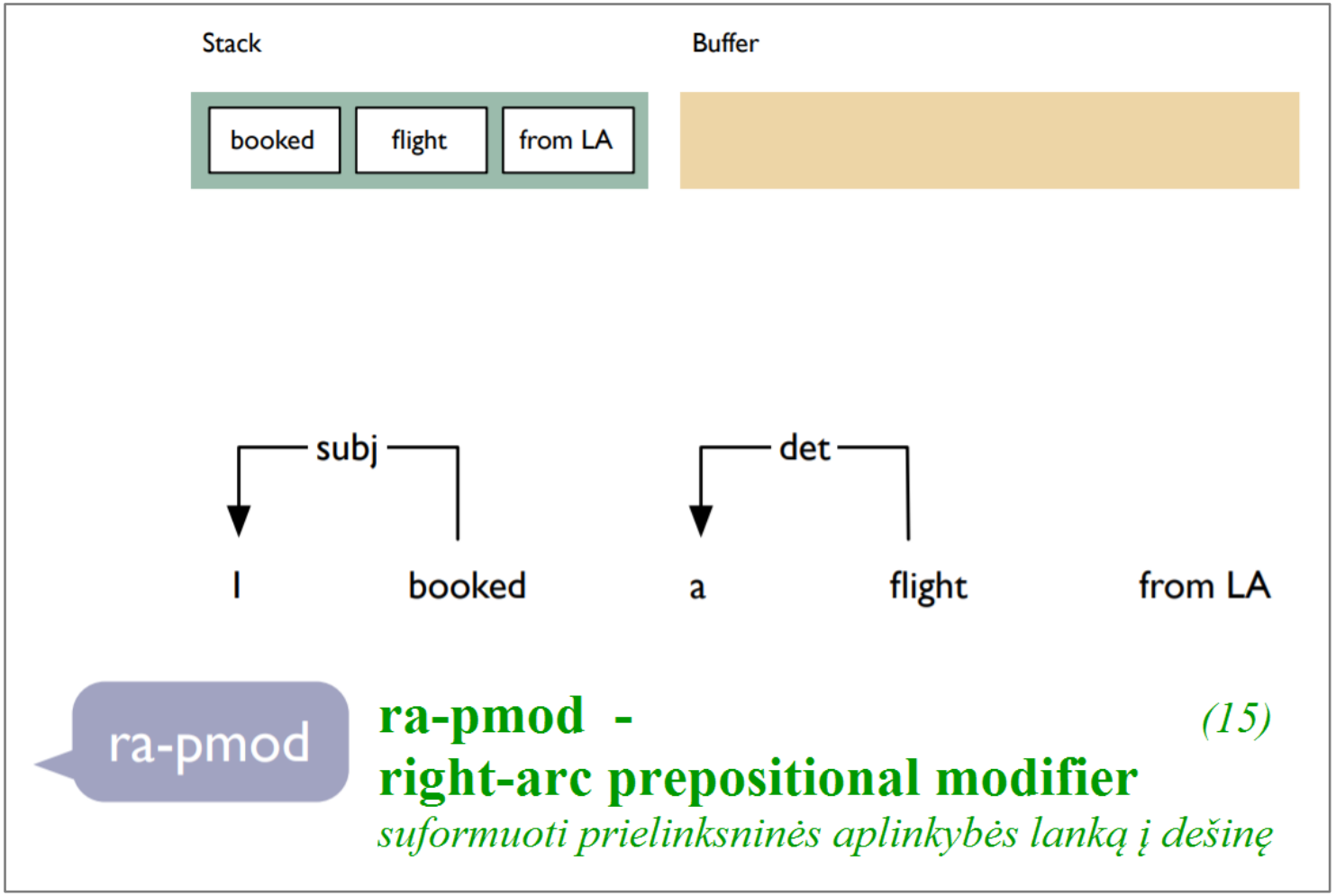
Suformuotas apibrėžiklio lankas į kairę ir kairysis juo susietas žodis "a" pašalintas iš steko (12)



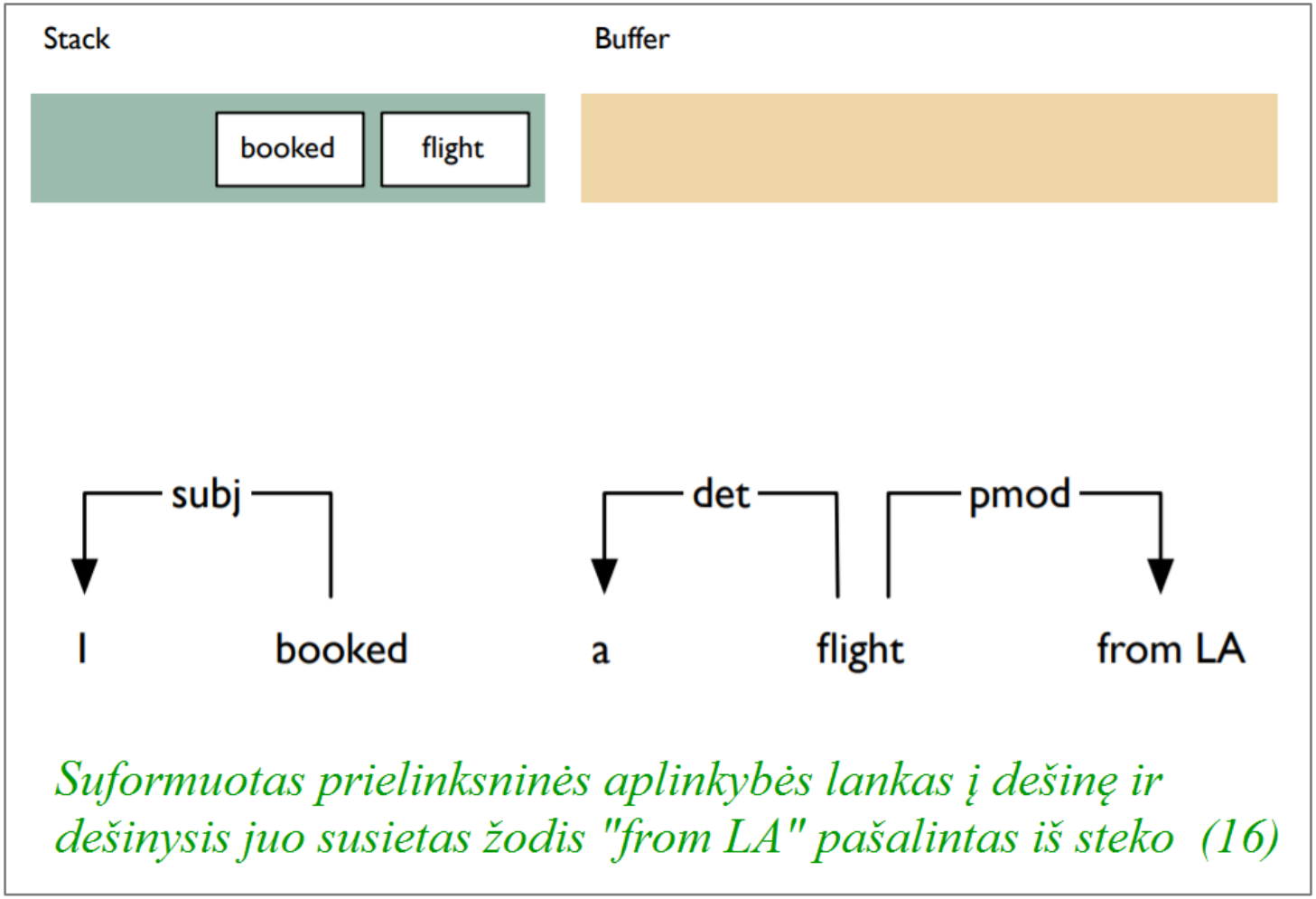
Kada ir kaip sudaryti žodžių junginius, sistema apmokoma naudojant auksinio standarto sakinius.



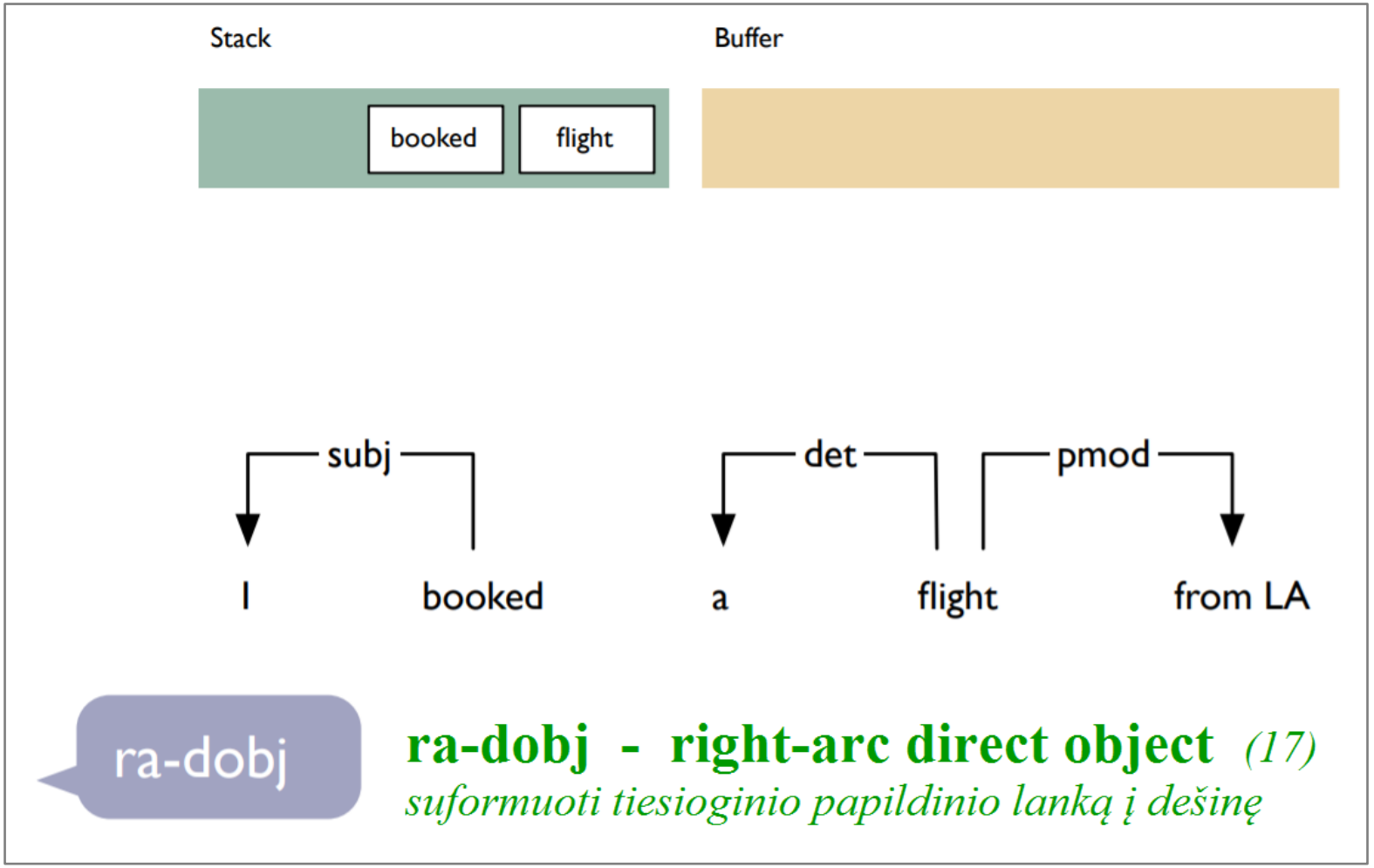
Žodis "from LA" perkeltas iš buferio į steką (14)

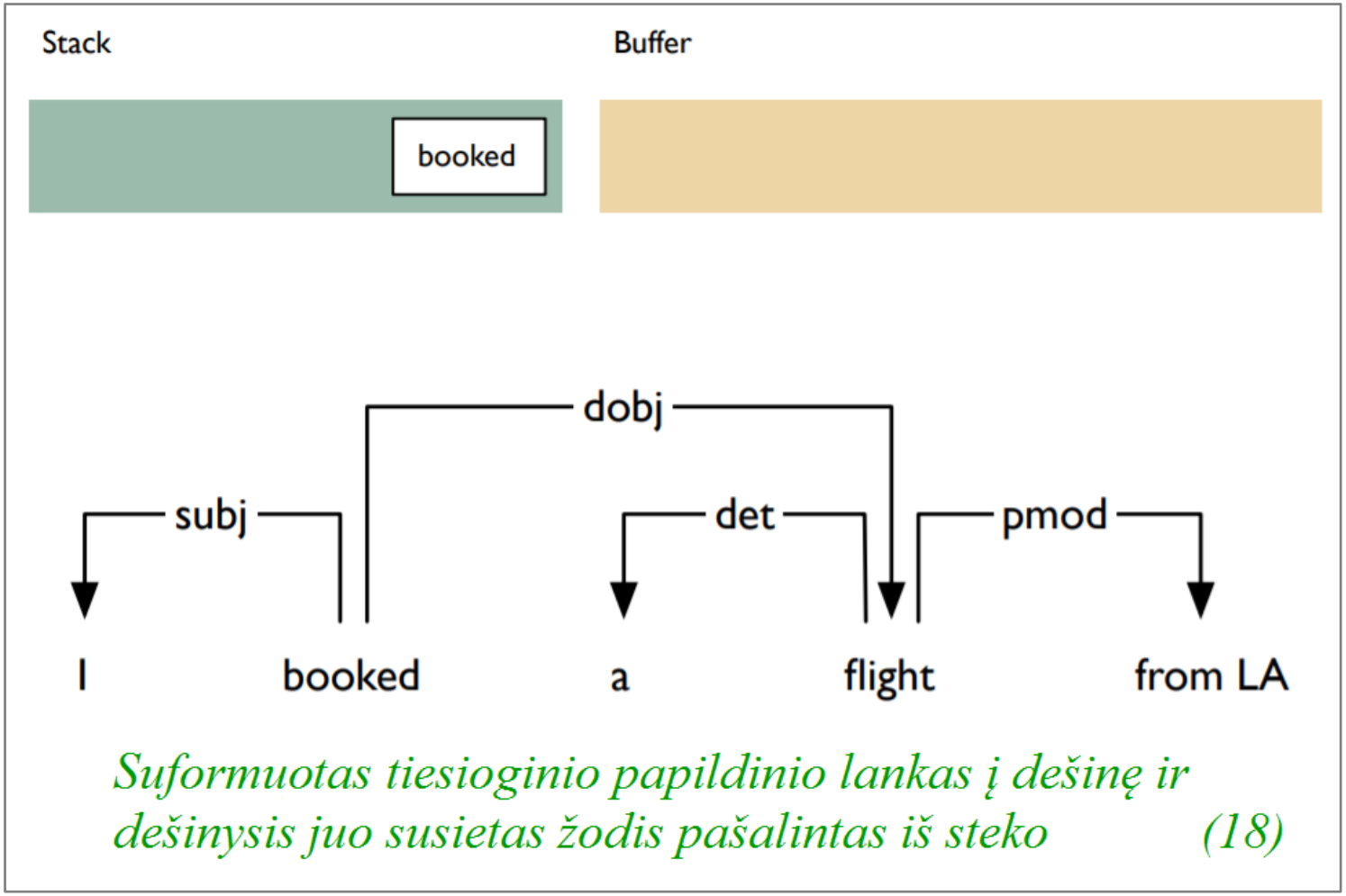


ra-pmod - *(15)*
right-arc prepositional modifier
suformuoti prielinksninės aplinkybės lanką į dešinę

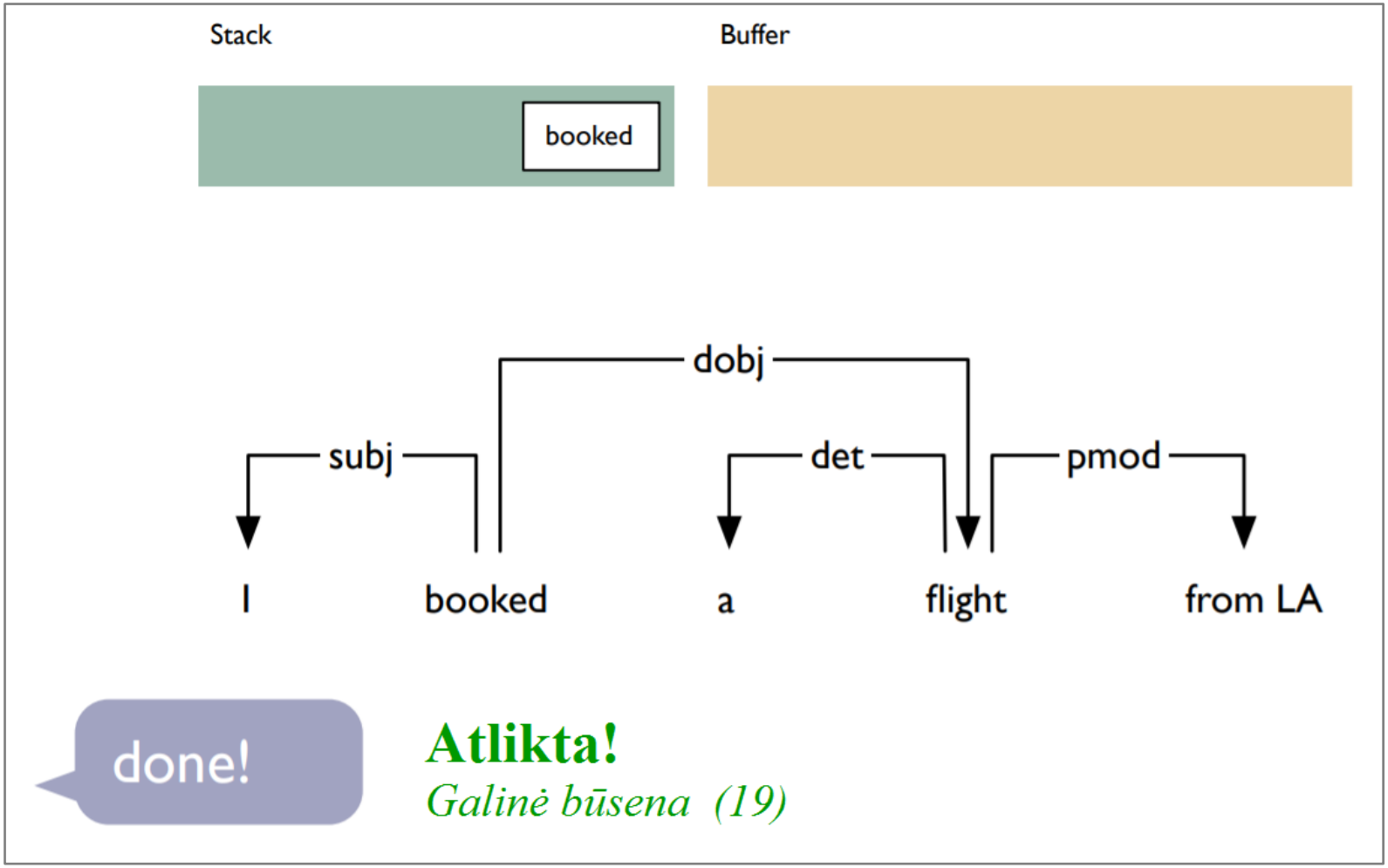


Suformuotas prielinksninės aplinkybės lankas į dešinę ir dešinysis juo susietas žodis "from LA" pašalintas iš steko (16)


Stack
Buffer
booked
flight
subj
det
pmod
I
booked
a
flight
from LA
ra-dobj
ra-dobj - right-arc direct object (17)
suformuoti tiesioginio papildinio lanką į dešinę



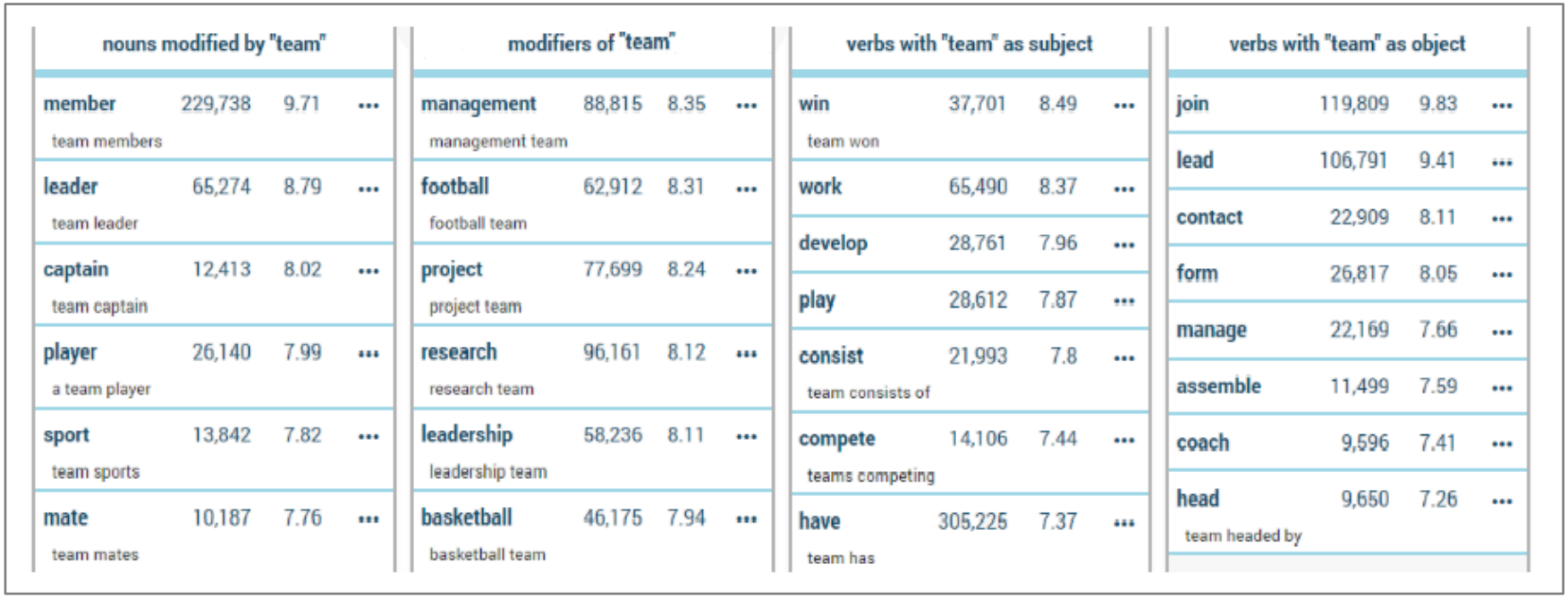
Suformuotas tiesioginio papildinio lankas į dešinę ir dešinysis juo susietas žodis pašalintas iš steko (18)



Atlikta!

Galinė būsena (19)

8 PRIEDAS: Žodžio *team* apžvalga



nouns modified by "team"			
member team members	229,738	9.71	•••
leader team leader	65,274	8.79	•••
captain team captain	12,413	8.02	•••
player a team player	26,140	7.99	•••
sport team sports	13,842	7.82	•••
mate team mates	10,187	7.76	•••

modifiers of "team"			
management management team	88,815	8.35	•••
football football team	62,912	8.31	•••
project project team	77,699	8.24	•••
research research team	96,161	8.12	•••
leadership leadership team	58,236	8.11	•••
basketball basketball team	46,175	7.94	•••

verbs with "team" as subject			
win team won	37,701	8.49	•••
work	65,490	8.37	•••
develop	28,761	7.96	•••
play	28,612	7.87	•••
consist team consists of	21,993	7.8	•••
compete teams competing	14,106	7.44	•••
have team has	305,225	7.37	•••

verbs with "team" as object			
join	119,809	9.83	•••
lead	106,791	9.41	•••
contact	22,909	8.11	•••
form	26,817	8.05	•••
manage	22,169	7.66	•••
assemble	11,499	7.59	•••
coach	9,596	7.41	•••
head team headed by	9,650	7.26	•••

9 PRIEDAS: *Lietuvių kalbos apžvalginė gramatika*: dvejopi ryšiai

has_adj_modifier/is_adj_modifier_of

has_right_modifier/is_right_modifier_of

has_adv_modifier/is_adv_modifier_of

has_noun_modifier/is_noun_modifier_of

has_acc_noun/is_acc_noun_of

has_nom_noun/is_nom_noun_of

has_ins_noun/is_ins_noun_of

has_gen_noun/is_gen_noun_of

has_dat_noun/is_dat_noun_of

has_loc_noun/is_loc_noun_of

has_inf_compl/is_inf_compl_of

has_pred_adj/is_pred_adj_of

has_adp/is_adp_of

has_conj/is_conj_of

10 PRIEDAS: Būsimojo laiko dalyvių pavartojimo internete pavyzdžiai

1. Daugiau už vaikus darželiuose **mokėsiantys** vilniečiai nustebo, kam bus naudojami jų pinigai https://www.15min.lt/naujiena/aktualu/lietuva/daugiau-uz-vaikus-darzeliuose-mokesiantys-vilnieciai-nustebo-kam-bus-naudojami-ju-pinigai-56-1257934?utm_source=rssfeed_front_main_v3&utm_medium=rss&utm_campaign=rssfeedgismeteo&copied [žiūrėta 2022-11-22]

2. Jau beveik dešimt metų jis treniruoja ir Tokijo olimpiadoje **startuosiančią** ir šiuo metu 12-ąją vietą pasaulio „Laser Radial“ moterų reitinguose užimančią buriuotoją Viktoriją Andrulytę. https://www.15min.lt/max/naujiena/gyvenimas/metu-treneris-l-eidukevicius-darbas-moteru-kolektyve-specifinis-tai-nera-vyriska-psichologija-1222-1256872?utm_source=rssfeed_front_main_v3&utm_medium=rss&utm_campaign=rssfeedgismeteo&copied [žiūrėta 2022-11-22]

3. Dabar **paspartėsiančios** tendencijos jau senokai ne naujiena. https://www.delfi.lt/news/daily/world/po-koronaviruso-pandemijos-atsiveres-pasaulis-atrodys-kitaip-prognozuoja-bent-tris-pokycius.d?id=84126391v [žiūrėta 2021-12-02]

4. Kitą savaitę **įsigaliosiantys** pakeitimai atliepia Lietuvos vairuotojų lūkesčius, nes didžioji dalis žmonių šį dokumentą norėtų gauti elektroniniu paštu. https://www.lrytas.lt/auto/saugus-eismas/2020/06/25/news/vairuotojams-aktualus-pakeitimai-jau-kita-savaite-ka-zinoti-del-svarbaus-dokumento-15395114/ [žiūrėta 2022-11-22]

5. Prie „Rergitros“ nusidriekė kilometrinės eilės: vairuotojai suskubo išvengti **įsigaliosiančio** mokesčio.
   https://www.delfi.lt/auto/autonaujienos/prie-regitros-nusidrieke-kilometrines-eiles-vairuotojai-suskubo-isvengti-isigaliosiancio-mokescio.d?id=84643931 [žiūrėta 2022-11-22]

6. Nors apie tikslų išleisiamų dainų kiekį R. Čivilytė kalba dar nedrąsia, ji dalinasi, kad, baigiantis rudeniui, tikisi **galėsianti** koncertuose atlikti ne vieną singlą iš būsimo albumo.
   https://www.lrytas.lt/zmones/muzika/2020/06/29/news/ivertinkite-pokycius-kardinaliai-stiliu-pakeitusi-rosita-civilyte-pristate-daina-man-gana--15441488/ [žiūrėta 2021-12-02]

7. R. Semeniukas pristatė naują vaizdo klipą: pokalbis – apie **užgimsiantį** naują albumą ir bliuzą Lietuvoje
   https://www.15min.lt/max/video/rsemeniukas-pristate-nauja-vaizdo-klipa-pokalbis-apie-uzgimsianti-nauja-albuma-ir-bliuza-lietuvoje-183782 [žiūrėta 2021-12-02]

8. Kabinos, prabėgus 60-iai metų, buvo gerokai aprūdijusios, pasenusios, o jų laukiančios stotys – apleistos ir atrodančios kaip tuoj **sugriūsiančios**.
   https://www.15min.lt/pasaulis-kiseneje/naujiena/kelioniu-pulsas/stalino-laiku-ciaturos-keltuvai-kurie-vadinti-skraidanciais-karstais-637-1349812?utm_source=rssfeed_front_main_v3&utm_medium=rss&utm_campaign=rssfeedgismeteo [žiūrėta 2022-11-22]

9. Šiandien greitai aštuonerių metų **sulauksiančio** šuniuko sveikata yra puiki.
   https://www.lrytas.lt/augintinis/prieziura/2020/11/15/news/zaibiskai-blogejant-suns-sveikatai-pasiule-ji-uzmigdyti-bet-kristina-isgelbejo-ji-is-mirties-17072543/ [žiūrėta 2022-11-22]

10. Detalus planas, kas keisis nuo trečiadienio: svarbiausi pokyčiai, **liesiantys** visus https://www.delfi.lt/news/daily/lithuania/detalus-planas-kas-keisis-nuo-treciadienio-svarbiausi-pokyciai-liesiantys-visus.d?id=86000125 [žiūrėta 2022-11-22]

11. Šiemet perlines vestuves **švęsiantys** dainininkai šią datą iki šiol dažniausiai paminėdavo pas giminaičius Vokietijoje, mat ten beveik prieš 30 metų juodu vyko medaus mėnesio. https://www.lrytas.lt/stilius/karamele/2021/02/16/news/kazlauskai-parode-nematytus-jaunystes-kadrus-ir-prisimine-meiles-istorija-viskas-vyko-tarsi-fime-18259232/ [žiūrėta 2022-11-22]

12. Šį vakarą **pasirodysiančioje** laidoje garsusis atlikėjų duetas pasakojo apie pažintį, kai Petras buvo jau pripažintas dainininkas, o Liveta – skurdžiai gyvenanti vieniša, jauna mama. https://www.lrytas.lt/zmones/tv-antena/2020/02/14/news/liveta-kazlauskiene-papasakojo-ka-isgyveno-kai-slapta-buvo-paviesintos-jos-nuogos-nuotraukos-13639800/ [žiūrėta 2022-11-22]

13. Armėnijos premjeras N. Pašinianas pranešė balandį **atsistatydinsiantis** https://www.15min.lt/naujiena/aktualu/pasaulis/armenijos-premjeras-n-pasinianas-pranese-balandi-atsistatydinsiantis-57-1477868?utm_source=gismeteo.lt&utm_medium=referral&utm_campaign=partnership&utm_content=all_right_300x802 [žiūrėta 2022-11-22]

14. Šįvakar „Euroviziją“ **atidarysiantys** „The Roop“ išlydėti į areną: prie viešbučio laukė užsienio žiniasklaida ir gerbėjai https://www.delfi.lt/veidai/eurovizija/sivakar-eurovizija-atidarysiantys-the-roop-islydeti-i-arena-prie-viesbucio-lauke-uzsienio-ziniasklaida-ir-gerbejai.d?id=87222885 [žiūrėta 2022-11-22]

15. To dar nebuvę: pasaulio čempionate **žaisiančiai** komandai treneris vadovaus sėdėdamas namuose https://www.15min.lt/sportas/naujiena/ziemos-sportas/to-dar-nebuve-pasaulio-cempionate-zaisianciai-komandai-treneris-vadovaus-sededamas-namuose-295-1505890?utm_source=gismeteo.lt&utm_medium=referral&utm_campaign=partnership&utm_content=all_right_300x802 [žiūrėta 2022-11-22]

16. Daugiau įrankių ir išteklių tėvams, **padėsiančių** tvarkyti funkcijas (įskaitant tai, kaip leisti vaikui iki 14 metų naudotis Paslauga ir „YouTube Kids“), rasite „YouTube“ Pagalbos centre arba „Google” „Family Link“. https://www.youtube.com/t/terms [žiūrėta 2022-11-22]

17. 5 daugiausia abejonių keliantys patarimai iš sodininkystės srities, ne tiktai **nepadėsiantys** išspręsti esamų problemų, bet dargi **lemsiantys** naujų atsiradimą. https://www.delfi.lt/agro/ukio-praktika-patarimai/5-internetiniai-patarimai-sodininkams-kuriais-patikejus-laukia-vien-nuostoliai.d?id=87282409 [žiūrėta 2022-11-22]

18. Atskleisti Ukrainos rinktinės marškinėliai, kuriuos šios šalies futbolininkai vilkės jau penktadienį **prasidėsiančiame** Europos čempionate, sukėlė emocijų audrą Rusijoje. https://www.15min.lt/sportas/naujiena/futbolas/ukrainieciu-marskineliai-isiutino-rusija-paragino-uefa-juos-uzdrausti-24-1516020?utm_source=gismeteo.lt&utm_medium=referral&utm_campaign=partnership&utm_content=all_right_300x802 [žiūrėta 2022-11-22]

19. Indrė Stonkuvienė griauna stereotipus, kad „špagatui“ yra per sena: nufilmuotoje pamokoje – jos patarimai **mesiantiems** sau šį iššūkį https://www.delfi.lt/veidai/zmones/indre-stonkuviene-griauna-stereotipus-kad-spagatui-yra-per-sena-nufilmuotoje-pamokoje-jos-patarimai-mesiantiems-sau-si-issuki.d?id=87707669 [žiūrėta 2022-11-22]

20. Už mylimojo Lauryno Suodaičio netrukus **ištekėsianti** Viktorija Siegel atšventė savo mergvakarį.
    Verslininkė, nuomonės formuotoja Viktorija Siegel, netrukus **tapsianti** mylimojo Lauryno Suodaičio žmona, su bičiulėmis trankiai atšventė savo mergvakarį
    https://www.delfi.lt/veidai/zmones/uz-mylimojo-lauryno-suodaicio-netrukus-istekesianti-viktorija-siegel-atsvente-savo-mergvakari.d?id=87674863
    [žiūrėta 2022-11-22]

21. Netrukus vestuves **atšoksianti** V. Siegel į lenktynes atvyko su L. Suodaičiu: „Šie metai laimingiausi mano gyvenime“
    https://www.lrytas.lt/zmones/veidai-ir-vardai/2021/07/17/news/netrukus-vestuves-atsoksianti-v-siegel-i-lenktynes-atvyko-su-l-suodaiciu-sie-metai-laimingiausi-mano-gyvenime--20125513/ [žiūrėta 2022-11-22]

22. Taivanas antradienį paskelbė Lietuvoje **atidarysiantis** atstovybę, kurios pavadinime bus vartojamas neoficialus šios savarankiškos salos pavadinimas, nors šis diplomatinis žingsnis neabejotinai suerzins Kiniją.
    https://www.delfi.lt/news/daily/world/po-zinios-apie-taivano-atstovybes-atidaryma-kinijos-ispejimas-lietuvai.d?id=87745533 [žiūrėta 2022-11-22]

23. E. Purauskytė pareiškė **nesinaudosianti** nepasiskiepijusiųjų paslaugomis
    https://www.15min.lt/vardai/naujiena/lietuva/e-purauskyte-pareiske-kad-nesinaudos-nepasiskiepijusiuju-paslaugomis-1050-1538588
    [žiūrėta 2022-11-22]

24. Kaip išsirinkti tinkamą pušų veislę, **džiuginsiančią** daugelį metų
    https://www.delfi.lt/gyvenimas/namai/kaip-issirinkti-tinkama-pusu-veisle-dziuginsiancia-daugeli-metu.d?id=87763587 [žiūrėta 2022-11-22]

25. Pirmadienio vakarą Rūdninkų poligone įrengtoje stovykloje migrantai surengė maištą – atvykusius žurnalistus pasitiko šūksniais apie žmogaus teises, laisvę, pranešė **badausiantys**, **negersiantys** vandens.
https://www.lrytas.lt/lietuvosdiena/aktualijos/2021/08/04/news/prakalbo-kokiu-migrantu-veiksmu-reiketu-tiketis-po-vrm-sprendimo-grazinti-i-baltarusija-20308960/ [žiūrėta 2022-11-22]

26. Prie vartų rikiuojasi pareigūnai, greičiausiai **bandysiantys** nustumti minią dar toliau.
https://www.15min.lt/naujiena/aktualu/lietuva/prie-seimo-renkasi-skiepu-priesininkai-nenori-galimybiu-paso-56-1547240?utm_source=gismeteo.lt&utm_medium=referral&utm_campaign=partnership&utm_content=all_right_300x802& [žiūrėta 2022-11-22]

27. **Dirbsiančiam** šį neįprastą darbą siūloma nelabai didelė alga, tačiau sunkiai plušėti neteks.
https://www.lrytas.lt/verslas/mano-pinigai/2021/08/31/news/auga-susidomejimas-nauja-lietuvoje-atsiradusia-profesija-sunkiai-pluseti-neteks-o-i-rankas-galima-gauti-iki-800-euru-20592776/ [žiūrėta 2022-11-22]

28. Pirmąjį filmą kosmose **kursianti** Rusijos komanda pripažinta tinkama skrydžiui
https://www.lrytas.lt/it/visata/2021/08/31/news/pirmaji-filma-kosmose-kursianti-rusijos-komanda-pripazinta-tinkama-skrydziui-20598947
[žiūrėta 2022-11-22]

29. Išrinkta nauja Lietuvos vasaros sostinė: nuo įspūdingo rekordinio labirinto iki **pakerėsiančių** nakvynės vietų
https://www.lrytas.lt/gamta/keliones/2021/09/01/news/isrinkta-nauja-lietuvos-vasaros-sostine-nuo-ispudingo-rekordinio-labirinto-iki-pakeresianciu-nakvynes-vietu-20604885 [žiūrėta 2022-11-22]

30. Bene didžiausio dėmesio „Sostinės dienose“ sulauks Katedros širdyje **koncertuosiantys** grupė iš Vokietijos „De-Phazz“, ekstravagantiškas elektroninės muzikos duetas „Beissoul & Einius“ bei rytojaus vakarą išskirtinį koncertą „Draugams“ dovanojantis Gytis Paškevičius su grupe. https://www.lrytas.lt/kultura/meno-pulsas/2021/09/03/news/prasideda-didziausias-vilniaus-miesto-festivalis-sostines-dienos-2021-kur-verta-nueiti-20631237 [žiūrėta 2022-11-22]

31. Jau visai netrukus Maljorkoje su mylimuoju Edgaru Eidėjumi (30) **susituoksianti** dainininkė, verslininkė Natalija Bunkė (38) spėja ne tik būti mama, rūpintis savo e-parduotuvės reikalais, aktyviai koncertuoti, bet ir diriguoti šventės organizatoriams. https://www.delfi.lt/veidai/zmones/vestuvems-besiruosianti-natalija-bunke-apie-pavardes-keitima-ceremonijai-kuriama-nekuklia-suknele-ir-jaunesnio-mylimojo-privalumus.d?id=88109625 [žiūrėta 2022-11-22]

32. Niekur nepradings ir visai valstybei toliau **kenksianti** valdančiųjų ir prezidento G. Nausėdos priešprieša. https://www.lrytas.lt/lietuvosdiena/aktualijos/2021/09/11/news/vytautas-bruveris-skaudus-smugis-r-karbauskiui-ir-nerima-valdantiesiems-keliantis-g-nausedos-elgesys-20715265 [žiūrėta 2022-11-22]

33. Trečiadienį posėdžiavęs Lietuvos valstiečių ir žaliųjų sąjungos suformuotas šešėlinis ministrų kabinetas pritarė nutarimui, kuriame ypač neigiamai įvertinti Ingridos Šimonytės Vyriausybės sprendimai, esą **bloginsiantys** regionų padėtį. https://www.delfi.lt/news/daily/lithuania/karbauskis-si-valdzia-suzlugdys-ir-regionus.d?id=88252625 [žiūrėta 2022-11-22]

34. Anot afganistaniečių, moteris vardu Laura jiems pasakė **susisieksianti** su „ambasada ar kažkuo“.
https://www.15min.lt/naujiena/aktualu/lietuva/lietuvoje-prieglobscio-prasiusiu-afganistanieciu-praeityje-rusijos-pedsakai-56-1569596?utm_source=gismeteo.lt&utm_medium=referral&utm_campaign=partnership&utm_content=all_right_300x802_main_v3 [žiūrėta 2022-11-22]

35. Prie oro uosto mane pasitikę ir kelionės metu **lydėsiantys** asmenys rankose jau turėjo man išduotą leidimą fotografuoti.
https://www.15min.lt/pasaulis-kiseneje/naujiena/kelioniu-istorijos/viena-reciausiai-lankomu-afrikos-genciu-iki-dantu-ginkluoti-ir-slapime-besimaudantys-mundari-vyrai-639-1575246?utm_source=gismeteo.lt&utm_medium=referral&utm_campaign=partnership&utm_content=all_right_300x802_main_v3 [žiūrėta 2022-11-22]

DAIVA ŠVEIKAUSKIENĖ

LIETUVIŲ KALBOS GRAMATIKOS KOMPIUTERIZAVIMAS

Mokslo studija

Redagavo *Daiva Šveikauskienė*, *Irutė Raišutienė, Teresė Paulauskytė*

Maketavo *Daiva Šveikauskienė*, *Rasa Labutienė*

Išleido Lietuvių kalbos institutas, P. Vileišio g. 5, LT-10308 Vilnius