ktu faculty of social sciences, arts and humanities 1922

46/2025
Research Journal
Studies about Languages
pp. 119–135
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
DOI 10.5755/j01.sal.1.46.40544

LINGUISTICS / KALBOTYRA

Naujų lietuvių kalbos anotuotų tekstynų rengimas: sandaros aspektai

Received 02/2025 | Accepted 06/2025

How to cite: Kovalevskaitė, J., Rimkutė, E., & Vaičenonienė, J. (2025). Naujų lietuvių kalbos anotuotų tekstynų rengimas: sandaros aspektai. Studies about Languages / Kalbų studijos, 46, 119–135. https://doi.org/10.5755/j01.sal.1.46.40544

Naujų lietuvių kalbos anotuotų tekstynų rengimas: sandaros aspektai

Developing new annotated corpora for Lithuanian: Compilation issues

JOLANTA KOVALEVSKAITĖ, Vytauto Didžiojo universitetas, Lietuva

ERIKA RIMKUTĖ, Vytauto Didžiojo universitetas, Lietuva

JURGITA VAIČENONIENĖ, Vytauto Didžiojo universitetas, Lietuva

Santrauka

Iki šiol parengti lietuvių kalbos gramatiškai anotuotų tekstynų ištekliai (morfologiškai anotuotas tekstynas MATAS, sintaksiškai anotuotas tekstynas ALKSNIS) yra nepakankamo dydžio atsižvelgiant į augančius lietuvių kalbos kompiuterizavimo poreikius. Todėl ES *NextGenerationEU* projekte „Morfologiškai ir sintaksiškai anotuotų tekstynų modeliai apmokymui (auksiniai standartai)“ yra rengiami du nauji tekstynai pagal tarptautinį *Universal Dependencies* (UD) standartą. Nauji tekstynai reprezentuos rašytinės lietuvių kalbos atmainą, todėl numatyta, kad juose, kaip ir „Dabartinės lietuvių kalbos tekstyne“ ar kituose anotuotuose tekstynuose, bus keturios dalys su tekstais iš grožinės, negrožinės (mokslinės), administracinės literatūros ir straipsniai iš internetinės periodikos.

Pristačius anotuotus tekstynus MATĄ ir ALKSNĮ, tarptautinio projekto UD kitų kalbų tekstynus, šiame straipsnyje aptariami naujų rengiamų anotuotų tekstynų sandaros aspektai: pirma, aprašoma planuojama tekstynų sandara ir proporcijos, antra, diskutuojamas pirminiam tekstyno automatinės analizės etapui svarbus klausimas – tekstyno skaidymas tekstyno vienetais (TV) (angl. *tokenization*) ir TV (angl. *token*) samprata. Nors tiek lietuvių, tiek kai kurių kitų kalbų tekstynų dydis gali būti nurodomas žodžiais, vis dėlto dažniausiai tekstynų dydis matuojamas TV, nes tekstynams aktualūs vienetai yra ne tik žodžiai, bet ir kiti elementai (skyrybos ženklai, skaitmenys, trumpiniai, simboliai). Kadangi TV samprata gali skirtis priklausomai nuo kalbos, tyrėjų sprendimo, svarbu, kad konkrečios kalbos automatinės analizės įrankyne būtų aprašyta, kokios TV skaidymo strategijos buvo laikomasi. Šiame straipsnyje paaiškiname, ką laikome TV, kaip traktuojami nevienareikšmiai skaidymo TV atvejai.

RAKTINIAI ŽODŽIAI: lietuvių kalba, automatinė morfologinė ir sintaksinė analizė, *Universal Dependencies* standartas, anotuoti tekstynai, tekstyno vienetas.

Įvadas

Lietuvių kalbos tekstynai, su kuriais prasidėjo ir tekstynų lingvistikos tyrimai Lietuvoje, pradėti rengti apie 1990-uosius, Vytauto Didžiojo universitete įkūrus Kompiuterinės lingvistikos centrą[1]. Pirmiausia 1992 m. pradėtas kaupti bendrasis rašytinės kalbos tekstynas „Dabartinės lietuvių kalbos tekstynas“[2].

Dabar SITTI interneto svetainėje[3] galima rasti keliasdešimt įvairaus pobūdžio tekstynų, duomenų bazių, žodynų, automatinės lietuvių kalbos analizės įrankių ir kitų išteklių. Išsamią informaciją apie šiuos išteklius (dažnai ir atvirą prieigą prie jų) galima rasti Bendrųjų kalbos išteklių ir technologijų infrastruktūros CLARIN-LT duomenų saugykloje[4].

Svarbi tekstynų grupė yra anotuoti tekstynai: šie tekstynai reikalingi automatinės kalbos analizės programoms rengti ir tobulinti, kitų anotuotų tekstynų tikslumui vertinti. Jie naudingi norint gauti kiekybinių duomenų apie kalbos dalių, gramatinių kategorijų, sintaksinių funkcijų ir kt. pasiskirstymą. Daug tyrimų atlikta automatinės morfologinės analizės srityje[5]: parengtas ir patobulintas morfologiškai anotuotas tekstynas MATAS[6], sukurti keli morfologiniai anotatoriai, kurių tikslumo rodikliai palyginti labai geri (Boizou et al., 2018, Kapočiūtė et al., 2017). Dėl viso šio automatinės morfologinės analizės įdirbio buvo galima vartotojams pristatyti automatiškai morfologiškai anotuotą „Dabartinės lietuvių kalbos tekstyno“ versiją, viešai prieinamą įrankį „Morfuoklis“[7], kuriuo galima ne tik anotuoti, bet ir susintezuoti (sugeneruoti) morfologines formas (plačiau apie analizės ir sintezės funkcijas žr. Rimkutė 2024; *Tour de CLARIN,* 2024[8]).

Kadangi po morfologinio anotavimo kitas automatinės teksto analizės etapas paprastai būna sintaksinė analizė, 2016 m. pradėtas rengti sintaksiškai anotuotas tekstynas ALKSNIS[9]. Deja, šis tekstynas negali patenkinti lietuvių kalbos kompiuterizavimo poreikių: jis per mažas, kad būtų galima vien remiantis šiuo tekstynu įgyvendinti tolesnius automatinės sintaksinės lietuvių kalbos analizės tikslus. Reikia daug didesnio sintaksiškai anotuoto tekstyno, kuris, paruoštas kaip auksinis standartas, būtų naudojamas sintaksinės analizės įrankiams tobulinti. Pagerėjus galimybėms automatiškai anotuoti didelius tekstų masyvus ir morfologiškai, ir sintaksiškai, galima būtų tobulinti įvairias lietuvių kalba grįstas kompiuterines technologijas, tokias kaip pokalbių robotai, didieji kalbos modeliai, dirbtiniam intelektui (toliau – DI) skirti ištekliai, su juo susiję įrankiai, o tai naudinga ne tik mokslinių tyrimų reikmėms, bet ir verslui, valstybinėms institucijoms – svarbu diegiant inovacijas viešojo valdymo, sveikatos, švietimo srityse.

Sintaksiškai anotuotų išteklių spragas padės užpildyti 2024–2026 m. vykdomas Europos Sąjungos *NextGenerationEU* projektas „Morfologiškai ir sintaksiškai anotuotų tekstynų modeliai apmokymui (auksiniai standartai)“[10]. Šiame projekte bus parengti 10 mln. žodžių morfologiškai ir sintaksiškai anotuoti lietuvių kalbos tekstynai.

[1] Kompiuterinės lingvistikos centras (KLC) nuo 2021 m. tapo naujai įsteigto Skaitmeninių išteklių ir tarpdisciplininių tyrimų instituto (SITTI) dalimi, institutui vadovauja prof. Rūta Petrauskaitė (Marcinkevičienė), tekstynų ir kompiuterinės lingvistikos pradininkė Lietuvoje.

[2] Šiuo metu prieinamos dvi šio tekstyno versijos: naujesnę, nuo 2012 m. prieinamą automatiškai morfologiškai anotuotą iš 208 mln. žodžių sudarytą versiją galima rasti http://corpus.vdu.lt/lt/, o senesnė, 2002 m. paskelbta, iš 140 mln. žodžių sudaryta neanotuota versija prieinama http://tekstynas.vdu.lt/tekstynas/.

[3] Žr. https://sitti.vdu.lt/.

[4] Žr. https://clarin.vdu.lt/xmlui/?locale-attribute=lt.

[5] VDU KLC, o dabar SITTI dėl šios srities kompetencijų yra vienas iš partnerių CLARIN tinklo žinių centre: CLARIN Knowledge Centre for Systems and Frameworks for Morphologically Rich Languages, https://www.kielipankki.fi/safmoril/.

[6] Naujausia šio tekstyno versija prieinama https://clarin.vdu.lt/xmlui/handle/20.500.11821/61.

[7] Žr. https://sitti.vdu.lt/morfuoklis/lt.

[8] Žr. https://www.clarin.eu/blog/tour-de-clarin-interview-erika-rimkute-and-virginijus-dadurkevicius.

[9] Žr. https://clarin.vdu.lt/xmlui/handle/20.500.11821/21.

[10] Projekto Nr. 02-098-K-0001; projektas finansuojamas Ekonomikos gaivinimo ir atsparumo didinimo plano „Naujos kartos Lietuva“ ir Lietuvos Respublikos valstybės biudžeto lėšomis, daugiau informacijos žr. https://sitti.vdu.lt/morfologiskai-ir-sintaksiskai-anotuotu-tekstynu-modeliai-dirbtinio-intelekto-apmokymui/.

Jie ypač reikalingi, kad būtų galima lietuvių kalbai pritaikyti inovatyvius technologinius sprendimus, įskaitant ir DI. Tokiai kalbai, kaip lietuvių, turime per mažai kalbinių duomenų DI poreikiams, todėl DI technologiją galime tobulinti ne duomenų kiekybe, o kokybe. Taigi labai reikalingi didesni morfologiškai ir sintaksiškai anotuoti tekstynai.

Pirmajame tekstynų rengimo etape – iki anotavimo – svarbu nutarti, kas bus laikoma tekstyno analizės vienetu, nes nuo to priklauso ir praktiniai dalykai – pradedant tekstyno dydžio apskaičiavimu ir baigiant tekstyno paruošimu anotavimui. Automatiškai apdorojant ir anotuojant tekstynus paprastai naudojamas vienetas, vadinamas *toukenu* (angl. *token*): tai natūraliosios kalbos analizės (angl. *natural language processing*) vienetas, apimantis ir žodžius, ir nežodžius (pvz., simbolius, skyrybos ženklus). Taigi paprastai tekstynų dydis nurodomas *toukenais*, nors tekstynų lingvistikoje dažnai naudojami ir lingvistinei analizei įprastesni vienetai – žodžiai. Iki šiol nėra nuoseklumo, kaip skaičiuojami lietuvių kalbos tekstynų dydžiai: pirmosios „Dabartinės lietuvių kalbos tekstyno“ versijos dydis nurodomas žodžiais, o antrosios – *toukenais;* ankstesnės MATO versijos aprašytos žodžiais, o paskutinė – *toukenais*, tiesa, nurodytas ir žodžių skaičius, plg.: 2024-12-19 į CLARIN-LT saugyklą įkeltą MATO versiją sudaro 2 137 287 *toukenai* (1 694 819 žodžiai (čia įeina ir skaitmenys)). Naujų rengiamų tekstynų dydis bus skaičiuojamas ir žodžiais, ir *toukenais,* pastarąją sąvoką toliau straipsnyje vadinsime lietuvišku terminu *tekstyno vienetas* (TV).

Taigi kitame skyriuje bus pristatyti iki šiol parengti gramatiškai anotuoti lietuvių kalbos tekstynai, aptarti jų sudarymo etapai, sandara. Rengiant didesnius anotuotus tekstynus, atsižvelgiama į kitų kalbų patirtį, todėl kitų kalbų anotuoti tekstynai apžvelgti antrame skyriuje. Trečias skyrius skirtas naujai rengiamų anotuotų tekstynų sandarai aptarti: atsižvelgiant į kitų kalbų taikomą praktiką, diskutuojama, kokių stilių ir žanrų tekstai, kokiomis proporcijomis juos turėtų sudaryti. Paskutinis poskyris skirtas TV problemai pristatyti ir paaiškinti, kas, rengiant naujuosius anotuotus tekstynus, bus laikoma TV, trumpai aprašyti TV nustatymo ir skaidymo gaires.

Gramatiškai anotuotų lietuvių kalbos tekstynų apžvalga

Šiame skyriuje bus apžvelgti du lietuvių kalbos gramatiškai anotuoti ir lingvistų sutvarkyti[11] tekstynai: morfologiškai anotuotas tekstynas MATAS ir sintaksiškai anotuotas tekstynas ALKSNIS. Kiek žinoma straipsnio autorėms, kol kas lietuvių kalbai nėra parengtas pakankamo dydžio, viešai prieinamas semantiškai anotuotas tekstynas. Vienas iš bandymų parengti eksperimentinį semantiškai anotuotą tekstyną (apie 20 tūkst. žodžių) buvo KLC, 2007–2008 m. vykdant projektą „Internetiniai ištekliai: anotuotas lietuvių kalbos tekstynas ir anotavimo priemonės (ALKA2)“. Šis tekstynas nėra viešai prieinamas.

Morfologiškai anotuoti tekstynai

Pirmuoju morfologiškai anotuotu lietuvių kalbos tekstynu galima laikyti Vidos Žilinskienės ir Laimos Grumadienės parengtą „Dabartinės rašomosios lietuvių kalbos dažninio žodyno elektroninį tekstyną“ (Grumadienė, 2002). Šis tekstynas sudarytas iš 1,2 mln. žodžių, jame esantys tekstai – originalūs (ne verstiniai), priklauso publicistiniam, moksliniam, beletristiniam[12] arba kanceliariniam stiliui. Iš šio tekstyno morfologiškai anotuotų[13] imčių (iš viso sudaryta 1200 imčių po 1000 žodžių pavartojimų) parengti šie žodynai: „Dabartinės rašomosios lietuvių kalbos dažninis žodynas (mažėjančio dažnio tvarka)“ (1997), „Dabartinės rašomosios lietuvių kalbos dažninis žodynas (abėcėlės tvarka)“ (1998) ir žodžių formų žodynas, turintis maždaug 150 000 žodžių formų, atspindintis skirtingų ir tų pačių žodžių vartojimą įvairiomis gramatinėmis formomis[14] (Grumadienė, 2002, p. 27–28). Šis tekstynas nebuvo ir iki šiol nėra viešai prieinamas.

[11] Tai reiškia, kad automatiškai suanotavus tekstus visas anotavimo žymas patikrina lingvistas (-ai) ir ištaiso pastebėtas automatinio anotavimo klaidas.

[12] Kituose tekstynuose šios dalies tekstai vadinami grožine literatūra.

[13] Naudota Vytauto Zinkevičiaus parengta MAN (morfologinio analizavimo ir formalizavimo) programa. Ją autorius pritaikė specialiai šiam darbui iš savo paties seniau parengtos kompiuterinės žodžių rašybos koregavimo programos, 2000 m. Vytauto Didžiojo universiteto užsakymu ją šiek tiek patobulino, perkėlė į aukštesnę versiją ir pavadino lemuokliu (Grumadienė, 2002, p. 27).

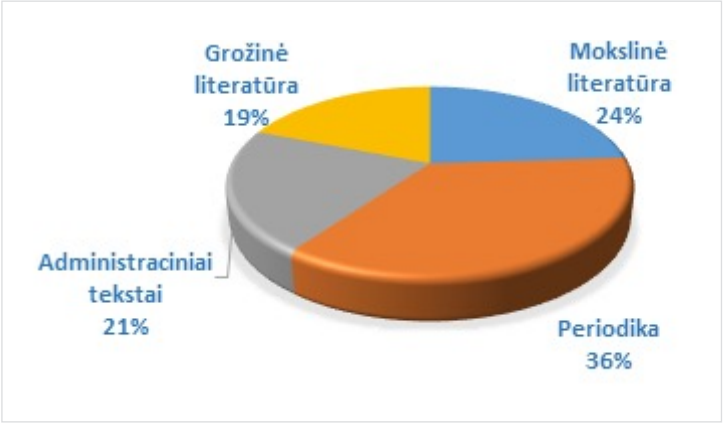
[14] Kiek žinoma, šis žodynas nepublikuotas ir viešai neprieinamas.

Morfologiškai anotuotas lietuvių kalbos tekstynas MATAS rengtas VDU KLC 2002–2014 m. Jo sudarymas vyko dviem etapais: 2000–2005 m. parengta pirmoji 1 mln. žodžių tekstyno versija, apie 2012–2014 m. šis tekstynas papildytas naujais tekstais, pertvarkytas. Pirmoji MATO versija[15] sudaryta pusiau automatiškai: tekstai anotuoti naudojant morfologinės analizės programą „Lemuoklis“ (Zinkevičius, 2000), vėliau visus žodžius peržiūrėjo ir rankomis sutvarkė E. Rimkutė. Tvarkant morfologiškai anotuotą tekstyną, susidurta su lietuvių kalbos morfologinio daugiareikšmiškumo problema: nustatyta, kad beveik pusė lietuvių kalbos kaitybinių formų yra daugiareikšmės (išsamiau žr. Rimkutė, 2006), todėl morfologiniame anotatoriuje buvo įdiegta vienareikšminimo funkcija (Rimkutė ir Daudaravičius, 2007). Pirmosios versijos pagrindu buvo parengtas „Dažninis rašytinės lietuvių kalbos žodynas“ (Utka, 2009), šio žodyno įvade pateiktas ir pirmąją versiją sudarančių tekstų sąrašas (iš viso 790 tekstų iš 35 šaltinių).

Antrosios papildytos, pertvarkytos MATO versijos MATAS v1.0[16] dydis – 1 693 410 žodžiai (144 047 sakiniai). Ši MATO versija anotuota „Lemuokliu“, bet šis tekstynas naudotas kitiems morfologiniams anotatoriams rengti: pavyzdžiui, kuriant ir tobulinant *semantika.lt* portale prieinamą automatinės morfologinės analizės programą (Dadurkevičius, 2017), taip pat naujausią morfologinės analizės ir sintezės įrankį „Morfuoklis“[17].

Tekstyne pateikti keturių funkcinių stilių tekstai: administracinio, grožinio, mokslinio ir periodikos, pagal stilių – publicistikos. Didžiausia yra publicistinio stiliaus dalis, ją sudaro 578 911 žodžių, toliau rikiuojasi mokslinio stiliaus dalis, ją sudaro 399 452 žodžiai, administracinio stiliaus dalį sudaro 335 848 žodžiai, mažiausia – grožinio stiliaus dalis, ją sudaro 327 086 žodžiai (procentinis pasiskirstymas matyti **1 paveiksle**).

Rengiant tiek anotuotus, tiek neanotuotus tekstynus, iškyla jų sandaros klausimas. Rengiant MATĄ, daugiau ar mažiau bandyta išlaikyti DLKT struktūrą, t. y. kad tekstyne būtų skirtingų stilių tekstų, ir DLKT tekstyno dalių proporcijas. Kaip matyti iš 1 paveikslo, visų tekstyno dalių, išskyrus periodiką, dydis yra panašus, o periodika tiek DLKT, tiek MATE sudaro didesnę dalį dėl šio stiliaus didesnės žanrinės, leksinės ir kitokios įvairovės. Galima teigti, kad išlaikyta ir stilistinė, ir žanrinė, ir teminė MATĄ sudarančių tekstų įvairovė. Nors šis tekstynas nedidelis, jo tekstų įvairovė pakankama (276 dokumentai, tiesa, skirtingo dydžio – nuo maždaug tūkstančio žodžių iki beveik šimto tūkstančių žodžių, bet tokių didelių tekstų nedaug), tekstai apima gana ilgą laikotarpį – 10–16 metų (1994–2011 metų tekstai). Kaip matyti iš trečiosios MATO versijos MATAS v3.0, tekstynas yra toliau tobulinamas: daugėja standartų, kuriais tekstynas pateikiamas (tekstynas buvo anotuotas keliais tarptautiniais standartais CoNLL-U, MULTEXT-East, taip pat lietuvišku standartu „Jablonskis“[19], plačiau apie skirtingų formatų MATĄ, žr. Bielinskienė et al., 2017, p. 3–5), tikslinamos anotavimo pažymos, tvarkomos techninės problemos (pvz., užsienio kalbų intarpai, simboliai ir kt.). Kadangi



1 pav. Morfologiškai anotuoto tekstyno MATAS sandara[18].

[15] Pirmoji MATO versija MATAS v0.2 prieinama internete: https://clarin.vdu.lt/xmlui/handle/20.500.11821/9.

[16] Prieinama: https://clarin.vdu.lt/xmlui/handle/20.500.11821/33.

[17] Žr. https://sitti.vdu.lt/morfuoklis/lt.

[18] Paveiksle matyti antrosios (MATAS v1.0) ir trečiosios (MATAS v3.0) versijos sandara. Trečioji MATAS v3.0 versija prieinama: https://clarin.vdu.lt/xmlui/handle/20.500.11821/61.

[19] Rimkutė E., Bielinskienė, A., Boizou, L., Utka, A., ir Dadurkevičius, V. (2024). *Morfologinio anotavimo standartas „Jablonskis“*. https://sitti.vdu.lt/jablonskis-lt/.

naudojamas kaip etalonas, šis tekstynas svarbus morfologinio anotavimo įrankiams tobulinti, kitų morfologiškai anotuotų tekstynų tikslumui įvertinti.

Kai kurie lietuvių kalbos tekstynai taip pat yra morfologiškai anotuoti, tačiau anotavimas atliktas tik automatiškai, lingvistų neperžiūrėtas, todėl juose pasitaiko anotavimo klaidų pvz., naujesnė DLKT versija[20], „Mokomasis lietuvių kalbos tekstynas“[21].

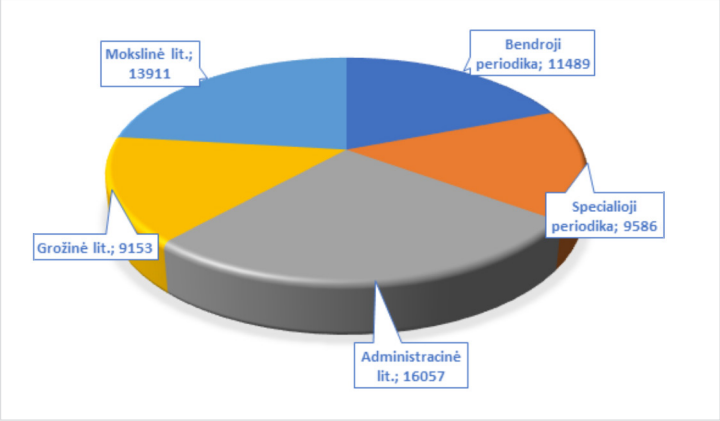
Sintaksiškai anotuotas tekstynas ALKSNIS

Viešai prieinamas vienintelis lietuvių kalbos sintaksiškai anotuotas lietuvių kalbos tekstynas ALKSNIS[22]. Jis pradėtas rengti KLC tyrėjų 2015–2019 m. kaip lietuvių kalbos sintaksinės analizės etalonas. Sintaksinė analizė apima tris lygmenis: automatinį žodžių ir sakinių segmentavimą bei morfologinę analizę[23], automatinę sintaksinę analizę ir rankinį automatinės (morfologinės ir sintaksinės) analizės tikrinimą (plačiau apie sintaksinę analizę žr. toliau).

Palyginti su morfologiškai anotuotu tekstynu, tekstynas ALKSNIS (v3.0) – dar labai mažas (3,6 tūkst. sakinių (60 196 žodžių)), nes jam parengti reikia skirti kelis kartus daugiau laiko nei morfologiškai anotuotam tekstynui. Be to, tekstynams, kurie rengiami kaip sintaksinės (ar kitokios) analizės etalonai, būdinga tai, kad jie būna nedideli, kiek įmanoma subalansuotos žanrinės sudėties, o sintaksinė analizė, nors ir atlikta automatiškai, bet patikrinta lingvistų.

Kaip teigia Bielinskienė et al. (2017, p. 27), vienas pagrindinių automatinės sintaksinės analizės tikslų – sukurti statistiniu pagrindu veikiantį sintaksinį analizatorių (angl. *statistically-based parser*), kuris būtų paremtas ALKSNIU kaip sintaksinės analizės etalonu ir būtų naudojamas dideliems tekstų kiekiams anotuoti. Tai iš dalies padaryta – ALKSNIS kaip mokymo išteklius naudojamas įrankyje UDPipe[24], t. y. su šiuo įrankiu galima automatiškai suanotuoti lietuvišką sakinį ar ilgesnę teksto ištrauką. Nors sintaksinės struktūros yra gana universalios, bet tiksliau anotuojama pasirinkus ALKSNIO tekstyno (o ne kitų kalbų sintaksiškai anotuotų tekstynų) modelį. Žinoma, reikia nepamiršti to, kad šis tekstynas gana mažas, todėl anotavimo tikslumas nėra didelis.

ALKSNIO v3.0 tekstyną sudaro keturios dalys (žr. **2 pav.**). Bendrosios bei specialiosios periodikos ir grožinės literatūros dalys yra apylygės, o administracinių tekstų dalis mažiausia, nes tai specifinė lietuvių kalba: sakiniai dažnai būna ilgi, juose gausu dokumentų pavadinimų, nuorodų į kitus dokumentus ir pan. Tekstynui sudaryti imti ištisi, nesutrumpinti tekstai. Grožinės literatūros dalį sudaro sakiniai iš lietuvių autorių kūrinių (apsakymų, novelių). ALKSNYJE naudoti tekstai, publikuoti 2004–2012 m. laikotarpiu (žr. Bielinskienė et al., 2016).



2 pav. ALKSNIS 3.0 sandara (pagal žodžių skaičių).

[20] Prieinama http://corpus.vdu.lt/lt/.

[21] Prieinamas https://kalbu.vdu.lt/mokymosi-priemones/mokomasis-tekstynas/.

[22] ALKSNIS prieinamas https://clarin.vdu.lt/xmlui/handle/20.500.11821/21. Šiuo metu CLARIN-LT saugykloje yra dvi versijos: 2017 m. versija ALKSNIS v2.1 (2 355 sakiniai), 2019 m. versija ALKSNIS v3.0. (3 643 sakiniai)

[23] Pirmoje ir antroje (ji žymima v3.0) ALKSNIO versijoje taikyti skirtingi morfologinio anotavimo standartai; pirmoje – MULTEXT-East, o antroje – „Jablonskio“ standartas.

[24] Žr. https://lindat.mff.cuni.cz/services/udpipe/.

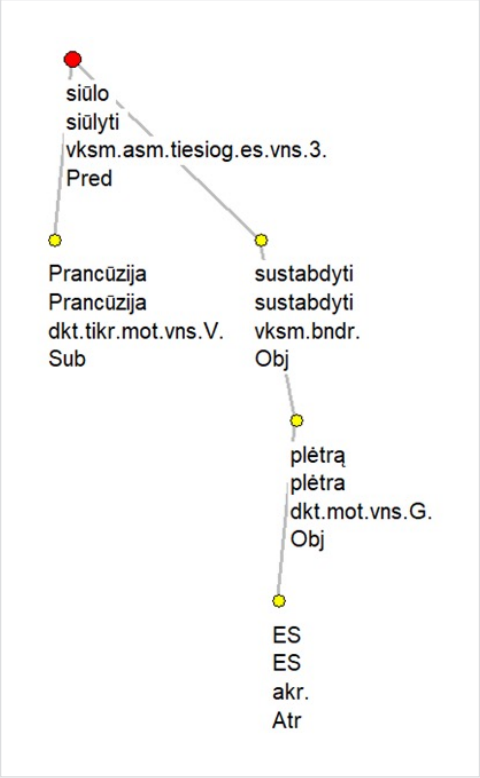
Tobulinant ALKSNĮ, jo v3.0 versijoje ne tik buvo pridėta naujų tekstų (mokslinių santraukų, dokumentų tekstų), patikslintos sintaksinės pažymos, pritaikytas „Jablonskio“ standartas, bet pažymėta ir tam tikra semantinio lygmens informacija – semantiškai nedalomi pastovieji junginiai (leksinės analitinės konstrukcijos ir frazeologizmai).

ALKSNIO tekstų automatinė sintaksinė analizė atlikta su VDU KLC mokslininkų sukurtu sintaksiniu analizatoriumi (plačiau žr. Boizou ir Zamblera, 2014). Analizatorius remiasi taisyklėmis pagrįstu (angl. *rule-based*) metodu, sukurtas *Haskell* kalbos pagrindu. Nuskaičius *semantika.lt* segmentavimo ir morfologinės analizės modulių morfologiškai anotuotus failus, pateikiami sakiniai su sugeneruotais priklausomybių medžiais (angl. *dependency trees*), kuriuose nurodomi sakinio dėmenų teminiai vaidmenys (angl. *thematic roles*) ir sintaksinės funkcijos (Boizou ir Zamblera 2014, p. 69). Atskiru įrankiu sintaksinės analizės rezultatai konvertuojami į PML (*Prague Markup Language*) formatą, kuris leidžia vizualizuoti ir redaguoti priklausomybių medžius (išsamiau apie pirmosios ALKSNIO versijos duomenų struktūrą, morfologinės ir sintaksinės analizės lygmenis, sintaksines funkcijas ir jų pažymas žr. Bielinskienė et al., 2017, p. 7–20).

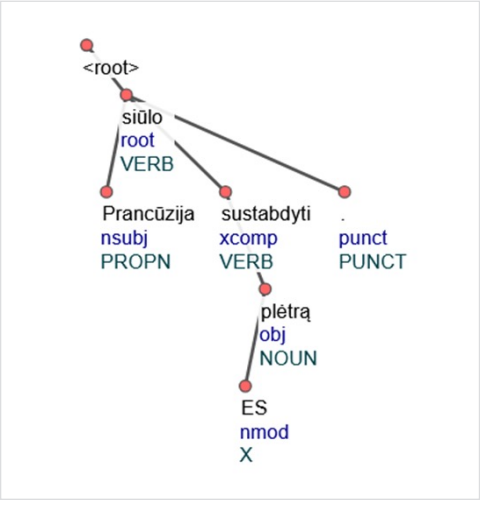
ALKSNIO sintaksinis anotavimas atliktas pagal priklausomybių gramatikos (angl. *dependency grammar*) modelį, šis modelis paprastai taikomas tipologiškai panašioms kalboms, kurioms būdinga žodžių formų kaityba ir laisva žodžių tvarka (naudotasi *Prague Dependency Treebank* (toliau – PDT) čekų modeliu, žr. Hajič et al., 2017). Tiesa, svarbu atkreipti dėmesį, kad ALKSNYJE yra tik morfologinis ir sintaksinis anotavimo lygmuo, o čekų sintaksiškai anotuotame tekstyne yra ir kitų lygmenų anotavimo, kur nurodomi diskurso ryšiai, pastovieji žodžių junginiai (plačiau Hajič et al., 2017).

Jau po pirmųjų ALKSNIO rengimo darbų buvo numatyta (Bielinskienė et al., 2017), kad šį tekstyną bus būtina konvertuoti į plačiai naudojamą tarptautinį *Universal Dependency* (toliau – UD) standartą (žr. Nivre, 2015; Zeman, 2018, naujausia apžvalga apie UD anotuotus tekstynus pateikta Marneffe et al., 2021). Tai naujausias ir dabar tarpkalbiniuose sintaksės tyrimuose ir automatinėje sintaksinėje analizėje plačiausiai naudojamas standartas. Parengus lietuvių kalbos išteklių šiuo standartu, sudaromos sąlygos aktyviau plėtoti minėtus lietuvių kalbos tyrimus. ALKSNIO v3.0 versija jau yra parengta ir UD standartu[25]. Naujai rengiamas sintaksiškai anotuotas 10 mln. žodžių tekstynas irgi bus anotuotas UD standartu.

Tiesa, PDT ir UD standartai gerokai skiriasi: pirmąjį iliustruoja **3a**, o antrąjį **3b** paveikslai, abiem standartais suanotuotas tas pats sakinys – *Prancūzija siūlo sustabdyti ES plėtrą*.



3a pav. Sintaksiškai anotuotas sakinys PDT standartu.



3b pav. Sintaksiškai anotuotas sakinys UD standartu.

[25] Ji prieinama UD tinklalapyje https://universaldependencies.org/.

Pirmiausia taikomas skirtingas sintaksinių pažymų skaičius, pvz., PDT standarte naudota apie 20 pagrindinių pažymų, o lietuviškus tekstus anotuojant UD standartu prireikia apie 40 pažymų (paminėtina, kad šios pažymos dar turi kombinacijų, t. y. sudarytos iš kelių pažymų, pirma pažymos dalis rodo pagrindinę pažymą, pvz., PDT standarte *Pred_Atr* pažyma reiškia, kad yra šalutinis dėmuo (tai rodo pažymos dalis *Pred*) ir tas šalutinis dėmuo yra atributinis (tai rodo pažymos dalis *Atr*); UD standarte pagrindinė subjekto pažyma yra *nsubj*, bet kai subjektas išreikštas neveikiamosios rūšies dalyviu, tada žymima *nsubj:pass*. Įprastai UD standarte objektai žymimi *obj*, bet jei tame pačiame sakinyje yra du nuo to paties žodžio priklausantys objektai, tai mažiau tiesiogiai veikiamas veiksmo (dažniausiai recipientas) objektas žymimas *iobj*. Apie UD standarto taikymą rengiant naują anotuotą tekstyną plačiau planuojama rašyti kitose publikacijose.

Rengiamas naujas sintaksiškai anotuotas 10 mln. žodžių tekstynas bus atskiras tekstynas, o ne naujesnė ALKSNIO versija. Naujojo tekstyno rengimo procese ALKSNIO v3.0 versija, anotuota pagal UD, bus naudojama sintaksinio įrankio mokymui ir anotavimo tikslumui įvertinti.

Kitų kalbų anotuotų tekstynų apžvalga

UD yra tarptautinis projektas, kuriame dalyvauja tyrėjai, rengiantys morfologiškai ir sintaksiškai anotuotus tekstynus. UD kaip anotavimo standartas šiuo metu jau yra pritaikytas daugiau nei šimtui kalbų, į projektą įsitraukę keli šimtai tyrėjų, projekto svetainėje skelbiama informacija apie 200 tekstynų, anotuotų UD standartu[26]. Rengiant šią apžvalgą, daugiausia remtasi UD projekto svetaine, kurioje galima rasti apibendrinančią informaciją apie sukurtus anotuotus tekstynus taikant UD anotavimo standartą: čia pateikiama susisteminta įvairių kalbų anotuotų tekstynų, jų sandaros, dydžių, matavimo vienetų ir kita informacija. Šiame rinkinyje yra sukaupti ir registruojami UD tekstynai, tačiau yra ir kitų tekstynų rinkinių, reprezentuojančių įvairesnius anotuotus tekstynus, pavyzdžiui, INESS[27], kur yra sukaupta informacija apie sintaksiškai anotuotus tekstynus pagal kalbas, skirtingus anotavimo metodus (angl. *constituency*, angl. *dependency*, angl. HPSG ir kt., plačiau apie skirtingus anotavimo metodus – Kübler ir Zinsmeister, 2015, p. 57–82), pagal skirtingas tekstynų kolekcijas (pvz., CLARIN-PL anotuotų tekstynų rinkinys, *Universal Dependencies 1.1* tekstynų rinkinys).

UD standartas susiformavo iš tipologinių gramatinių teorijų (angl. *typologically oriented grammatical theories*): čia gramatiniai santykiai tarp žodžių rodo, kaip skirtingose kalbose morfosintaksiškai žymimos predikatų ir argumentų struktūros, o informacija apie žodžius perteikiama nurodant morfologinius požymius ir kalbos dalį. Šis požiūris leidžia nuosekliai anotuoti tipologiškai skirtingas kalbas, taip plečiant automatinio natūraliosios kalbos apdorojimo ir lingvistinių tyrimų galimybes (plačiau žr. Marneffe et al., 2021). UD projekte siekiama pateikti universalią taksonomiją (anotavimo kategorijų sąrašą) ir gaires, kurias galima būtų taikyti ir plėsti atsižvelgiant į kiekvienos kalbos ypatumus, t. y. kad būtų galima ne tik nuosekliai anotuoti panašias konstrukcijas įvairiose kalbose nepaisant žodžių tvarkos, morfologijos ir kt. skirtumų, bet prireikus išplėsti taksonomiją konkrečioms kalboms reikalingomis kategorijomis (Marneffe et al., 2021, p. 256).

Kadangi rengiamas sintaksiškai anotuotas lietuvių kalbos tekstynas remiasi UD standartu, toliau šioje trumpoje apžvalgoje aptarsime būtent šio tipo anotuotus tekstynus: pirma, apžvelgsime esamus fleksinių baltų ir slavų kalbų tekstynus ir, antra, palyginsime anotuotų tekstynų sudarymo principus daug išteklių turinčioje anglų kalboje.

1-oje lentelėje matome, kad tekstynai labai įvairūs, gali apimti šnekamosios (retesni, nors jau parengti daugeliui kalbų (žr. apžvalgą Dobrovoljc, 2022), rašytinės arba abi kalbos atmainas; gali būti daugiažanriai arba sudaryti tik iš konkrečių žanrų tekstų. Vis dėlto labiau dominuoja naujienų, grožinės literatūros žanrai. Palyginamumą kartais apsunkina tai, kad tie patys žanrai įvardijami skirtingais terminais, pavyzdžiui, vieni tekstynai apibūdina tekstų pobūdį (laiškai, socialinių tinklų įrašai, komentarai, recenzijos), kiti tiesiog pamini medijų rūšį (socialinės medijos). Prie konkrečių tekstynų apie žanrų proporcijas ar balansą pateikiama nedaug informacijos.

[26] Žr. https://universaldependencies.org/introduction.html.

[27] Žr. https://clarino.uib.no/iness.

1 lentelė. Anotuotų baltų ir slavų kalbų UD standarto tekstynų dydžiai ir sandara.[28]

Kalba	Pavadinimas	Dydis	Sandara[29]
Baltarusių	*Belarusian HSE*	25231 sakinys 305109 TV	Grožinė literatūra, teisiniai tekstai, naujienos, negrožinė literatūra, socialinės medijos, poezija, *Vikipedija*
Bulgarų	*Bulgarian BTB*	11138 sakiniai 156149 TV	Naujienos, teisiniai tekstai, grožinė literatūra
Čekų	*Czech PDT*	87907 sakiniai 1,5 mln. TV	Naujienos, verslo, mokslo populiarinimo tekstai, apžvalgos, negrožinė literatūra
	Czech CAC	24709 sakiniai 493306 TV	Naujienos, negrožinė literatūra, teisiniai tekstai, apžvalgos, medicinos tekstai
	Czech CLTT	1121 sakinys 35997 TV	Teisiniai tekstai
	Czech FicTree	12760 sakinių 166432 TV	Grožinė literatūra
	Czech PUD	1000 sakinių 18565 TV	Naujienos, *Vikipedija*
Kroatų	*Croatian SET*	9010 sakinių 199409 TV	Naujienos, žiniatinklis, *Vikipedija*
Latvių	*Latvian-LVTB*	18850 sakinių 317369 TV	Naujienos, grožinė literatūra, teisiniai, akademiniai tekstai, šnekamoji kalba
Lenkų	*Polish PDB*	22152 sakiniai 350000 TV	Grožinė literatūra, negrožinė literatūra, naujienos
	Polish LFG	17246 sakiniai 130967 TV	Grožinė literatūra, negrožinė literatūra, naujienos, šnekamoji kalba, socialinės medijos
	Polish PUD	1000 sakinių 18333T V	Negrožinė literatūra, naujienos
Makedonų	*Macedonian MTB*	155 sakiniai 1360 TV	Pavyzdiniai gramatikos sakiniai
Slovakų	*Slovak SNK*	10604 sakiniai 106043 TV	Grožinė literatūra, negrožinė literatūra, naujienos
Slovėnų	*Slovenian SSJ*	13000 sakinių 267097 TV	Grožinė literatūra, negrožinė literatūra, *Vikipedija*, naujienos
	Slovenian SST	6104 sakiniai 76341 TV	Šnekamoji kalba
Serbų	*Serbian SET*	4384 sakiniai 97673 TV	Naujienos
Ukrainiečių	*Ukrainian IU*	7000 sakinių 122000 TV	Tinklaraščiai, el. laiškai, grožinė literatūra, naujienos, nuomonių straipsniai, *Vikipedija*, teisiniai dokumentai, laiškai, pranešimai ir komentarai

[28] Informacija imta iš https://universaldependencies.org/; vardijimas abėcėlės tvarka.

[29] Tekstynai nebūtinai sudaryti iš pilnų tekstų, tekstynus ar jų dalis gali sudaryti ne pilni tekstai, o tekstų ištraukos (plačiau žr. kitą poskyrį).

Slavų kalbų tekstynai pagal žanrus ir subalansavimą sunkiai palyginami. Taip pat skiriasi ir tekstynų dydžiai: mažiausią – makedonų – tekstyną sudaro 155 sakiniai, o didžiausią – čekų – tekstyną 87 907 sakiniai ir 1,5 mln. TV. Vidutinis tekstynų dydis apie 10 000 sakinių. Iš šių duomenų matyti, kad kuriamas 10 milijonų žodžių sintaksiškai anotuotas lietuvių kalbos tekstynas ženkliai viršytų ne tik mažiau išteklių, bet ir didžiųjų, išteklių turtingų kalbų anotuotų tekstynus.

Tačiau kol kas lietuvių kalbos tekstynas ALKSNIS (3 643 sakiniai ir 60 196 žodžiai) gerokai mažesnis ir už giminiškos latvių kalbos sintaksiškai anotuotą tekstyną[30]. Latvių kalbos tekstynas rengiamas nuo 2010 m., CLARIN.lv saugykloje įkeltos jau kelios šio tekstyno versijos, naujausia viešai prieinama versija yra v2.15, ją sudaro 19 367 sakiniai, 327 464 TV[31]. Latvių tekstynas pirmiausia parengtas taikant hibridinį anotavimo modelį *SemTi-Kamols* (angl. *hybrid dependency-constituency grammar model*), nuo 2016 m., prisijungus prie UD bendruomenės, pateikiama ir sintaksiškai anotuoto tekstyno versija UD standartu, angl. *Latvian UD Treebank* (Pretkalniņa et al., 2018).

Apibendrinant UD svetainėje pateiktą informaciją, galima teigti, kad daugiausia anotuotų tekstynų turi šios kalbos: anglų (10), italų (10), turkų (9), kinų (7), prancūzų (7). Skaičiuojant visų konkrečios kalbos anotuotų tekstynų dydžius, didžiausi rinkiniai siekia apie 1 (pvz., ispanų, portugalų, lotynų), 2 (pvz., japonų, islandų, čekų) ar 3 milijonus TV (pvz., vokiečių). Toliau palyginsime daug išteklių ir daugiausia tekstynų turinčios anglų kalbos tekstynų dydžius ir sandarą.

2 lentelė. Anotuotų anglų kalbos tekstynų dydžiai ir sandara.[32]

Pavadinimas	Dydis	Sandara
English Atis	5432 sakiniai 61879 TV	Negrožinė literatūra, naujienos
English ESLSpok	2320 sakinių 21312 TV	Šnekamoji kalba
English EWT	16622 sakiniai 251492 TV	Tinklaraščiai, socialiniai tinklai, apžvalgos, el. laiškai, žiniatinklis
English GENTLE	1334 sakiniai 17617 TV	Akademiniai tekstai, gramatikose pateikiami pavyzdžiai, teisiniai tekstai, medicininiai tekstai, negrožinė literatūra, poezija, socialinės medijos, šnekamoji kalba
English GUM	12147 sakiniai 208332 TV	Akademiniai tekstai, tinklaraščiai, el. laiškai, grožinė literatūra, administraciniai tekstai, teisiniai tekstai, naujienos, negrožinė literatūra, socialiniai tinklai, šnekamoji kalba, žiniatinklis, *Vikipedija*
English GUMReddit	895 sakiniai 15958 TV	Tinklaraščiai, socialinės medijos
English LinES	5243 sakiniai 93200 TV	Grožinė literatūra, negrožinė literatūra, šnekamoji kalba
English ParTUT	2090 sakinių 49602 TV	Teisiniai tekstai, naujienos, *Vikipedija*
English Pronouns	285 sakiniai 1640 TV	Gramatikose pateikiami pavyzdžiai
English PUD	1000 sakinių 21051 TV	Naujienos, *Vikipedija*

[30] Sintaksiškai anotuotas latvių kalbos tekstynas prieinamas https://sintakse.korpuss.lv/. Čia pateikiama informacija apie tekstyno sandarą, TV sampratą, morfologinio anotavimo principus.

[31] Rituma, L., et al. (2024). *LVTB - Latvian Treebank v2.15 (2024-11-15)*, CLARIN-LV digital library at IMCS, University of Latvia, http://hdl.handle.net/20.500.12574/112.

[32] Žr. https://universaldependencies.org/en/index.html

Nors morfologiškai (ir (ar) sintaksiškai) anotuotų anglų kalbos tekstynų yra ženkliai daugiau nei atskirų slavų ar baltų kalbų, bet dydžiu jie panašūs, varijuoja nuo kelių šimtų iki daugiau nei 10 000 sakinių. Žanrine įvairove pasižymi didžiausi tekstynai – *English EWT*, *English GENTLE* ir *English GUM*.

Iš prieinamų duomenų matyti, kad kitų kalbų tekstynus sudaro įprastai tekstynuose dažniausiai reprezentuojami rašytinės kalbos tekstai, tačiau jau nemažai ir tokių, kur įtraukta sakytinė kalba ir internetinės komunikacijos žanrai (el. laiškai, komentarai, tinklaraščiai, socialinės medijos ir pan.). Lietuvių kalbos tekstyne bus kaupiami tik rašytinės kalbos tekstai išlaikant ALKSNYJE pradėtą formuoti sandarą: grožinės, mokslinės, administracinės literatūros tekstai ir periodikos tekstai (straipsniai iš naujienų portalų). Didelė dalis internetinės komunikacijos taip pat vyksta raštu, todėl ateityje galbūt svarstytina parengti atskirą tekstyną arba rengiant sakytinės kalbos tekstyną kaip vieną iš dalių numatyti ir komunikaciją internete.

Tekstų atranka ir tekstyno analizės vienetai

Įvade minėtame projekte „Morfologiškai ir sintaksiškai anotuotų tekstynų modeliai apmokymui (auksiniai standartai)“ bus parengti du tekstynai: vienas morfologiškai anotuotas, o kitas bus anotuotas ir morfologiškai, ir sintaksiškai. Abu tekstynus sudarys tie patys tekstai, todėl šiame skyriuje bus aptariama tekstyno sandara aktuali abiem tekstynams ir aprašyta tekstyno dydžio skaičiavimo metodika ir TV sąvoka. Tiek morfologiškai, tiek sintaksiškai anotuotiems tekstynams svarbu nustatyti TV ribas ir juos atitinkamai anotuoti.

Naujų gramatiškai anotuotų tekstynų sandara

Iš 1 lentelės duomenų galima matyti, kad, jeigu vienai kalbai yra keli anotuoti tekstynai (kaip kad čekų atveju), daugiausia anotuotų tekstų yra iš naujienų (publicistikos) ir negrožinės literatūros, tada eina grožinė literatūra, o dokumentų tekstynas mažiausias. *Prague Dependency Treebank* sandara – pagrindinis tekstynas tik iš publicistikos tekstų (dienraščių ir žurnalų), bet PDT sintaksiškai anotuotų tekstynų rinkinyje (angl. *Prague Dependency Treebank-Consolidated* 1.0 (PDT-C 1.0)) jau yra tekstynų, sudarytų iš verstinės kalbos (*Wall Street Journal* versija čekų kalba) ir sakytinės čekų kalbos įrašų (Hajič et al., 2020).

Kaip minėta, rengiamuose lietuvių kalbos anotuotuose tekstynuose bus kaupiami tik rašytinės kalbos tekstai išlaikant ALKSNYJE numatytą sandarą: grožinės, mokslinės, administracinės literatūros ir publicistikos tekstai. Numatyta imti originalios rašytinės kalbos tekstus, nes vertimų kalba skiriasi nuo originalo kalbos tekstų (tačiau reikia turėti omenyje, kad, pavyzdžiui, publicistikos tekstuose (straipsniuose) bus ir vertimų).

Kalbant apie proporcijas, kaip ir kitų kalbų atveju, daugiausia anotuotų tekstų planuojama kaupti iš publicistikos (įvairiatemiai ir įvairiažanriai straipsniai, publikuojami naujienų portaluose), šie tekstai sudarys apie 50 proc. tekstyno. Dokumentai ir mokslinės literatūros tekstai sudarys po 20 proc., o grožinės literatūros tekstai – apie 10 proc. Skirtingų dalių proporcijos svarbios rengiant tekstyną kaip standartą: kuo daugiau reprezentatyvių duomenų bus sukaupta iš įvairių rašytinės kalbos stilių, tuo naudingesnis šis tekstynas bus kaip anotavimo standartas. Be to, naujuose tekstynuose bus tik nuo 2000 m. išleisti / parengti tekstai, nes naujesniais tekstais taip pat geriau reprezentuojama dabartinė kalba. Čia galima pridurti, kad tekstynus sudarys standartinę, bendrinę kalbą atspindintys tekstai, nors gali pasitaikyti ir nestandartinės kalbos atvejų: pvz., moksliniuose straipsniuose, kur dėl tyrimo specifikos yra pateikiama tarminio ar nenorminio kalbėjimo pavyzdžių, senesnės vartosenos pavyzdžių, taip pat grožinėje literatūroje perteikiant šnekamąją kalbą.

Tekstynų formavimas labai priklauso ne tik nuo numatytų proporcijų ir tekstų tipo, stiliaus, žanro, kuriuos norima atspindėti tekstynuose, bet ir galimybių gauti ir (ar) panaudoti reikiamus tekstus. Kai tekstynai rengiami tam, kad vėliau būtų panaudojami tik moksliniams tikslams, galimybių panaudoti publikuotus tekstus yra daugiau: visi tekstai neplatinami, per atvirą tekstynų paiešką prieinamos tik tam tikro ilgio citatos (pvz., „Dabartinės lietuvių kalbos tekstyne“, „Mokomajame lietuvių kalbos tekstyne“ prieinamos teksto ištraukos iki 600 simbolių). Kai siekiama parengti tokio tipo tekstyną, kuris kaip anotavimo standartas būtų prieinamas tiek mokslo, tiek komerciniams tikslams, Lietuvoje atsiranda daugiau apribojimų dėl tekstų naudojimo galimybių.

Administracinės literatūros tekstai yra neproblemiškiausi autorių teisių požiūriu. Kadangi jie paprastai yra viešai prieinami, naujuose tekstynuose numatyta pateikti gana didelę jų įvairovę: Lietuvos Respublikos Konstitucinio

Teismo nutarimų, Lietuvos Respublikos įstatymų, taip pat kitų tvarkomųjų (įsakymai, potvarkiai, sprendimai, protokolai), organizacinių (įstatai, statutai, taisyklės, planai), informacinių tekstų (ataskaitos, strategijos). Tiesa, nors šie tekstai yra lengviausiai prieinami, tačiau jų tekstynuose reikia gana nedidelio kiekio – dokumentų kalba yra specialiojo, o ne plačiai vartojamo diskurso dalis, todėl dokumentuose vartojamos sintaksinės struktūros neturi dominuoti tekstyne, kuris rengiamas kaip etalonas.

Mokslinio patekstynio dalis formuojama iš mokslinių straipsnių (imant atvirosios prieigos tekstus iš duomenų bazės „Lituanistika“) ir mokslinio stiliaus (neverstinių) knygų, leistų specializuotuose leidyklose, universitetuose. Iš neverstinių grožinės literatūros kūrinių (novelių, apsakymų, romanų) tikimasi suformuoti apie 1 mln. žodžių patekstynį. Tačiau reikia pripažinti, kad galutinės proporcijos bus aiškios baigus formuoti tekstyną ir išsprendus autorių teisių klausimus, nes, vadovaujantis dabar galiojančia autorių teisių tvarka, iš leidyklų ar knygų autorių būtina gauti sutikimą naudoti jų kūrinius ir (arba) tų kūrinių dalis, nes anotuoti tekstai atsiras atvirosios prieigos tekstynuose.

Atrenkant mokslinio ir grožinio stiliaus tekstus, atsižvelgiama į tai, kad ir vienoje, ir kitoje dalyje būtų kuo didesnė autorių, temų įvairovė: pavyzdžiui, grožinės literatūros dalis būtų įvairi, jeigu pavyktų įtraukti kuo daugiau skirtingų autorių, įvairių žanrų, temų tekstų, o mokslinės literatūros dalis būtų pakankamai reprezentatyvi, jeigu joje būtų straipsnių iš skirtingų disciplinų žurnalų; skirtingų autorių, skirtingų temų straipsnių (knygų); mokomosios, pažintinės literatūros tekstų. Tokiomis gairėmis vadovaujantis formuojama tekstynų sandara, nuo jos, kaip minėta, nemažai priklauso šių tekstynų-standartų panaudojimo galimybės įvairiems lietuvių kalbos automatinės analizės uždaviniams ir tų uždavinių sprendimo kokybė.

Rengiami gramatiškai anotuoti lietuvių kalbos tekstynai sandara panašūs į slovėnų ir latvių kalbų tekstynus, bet jų sudarymo principai yra kiek kitokie – žanrinę, autorių įvairovę taip pat siekiama užtikrinti, tačiau minėtų kalbų tekstynai sudaryti iš tekstų fragmentų, be to, priimti nevienodi sprendimai dėl balansavimo.

Kaip jau minėta ankstesniame skyriuje, latvių kalbos sintaksiškai anotuoto tekstyno *Latvian Treebank* v2.15 versiją sudaro 9 367 sakiniai, 327 464 TV. Rituma et al. (2019) aprašo principą, kaip buvo sudaromos imtys: reprezentatyvi maždaug 10 000 sakinių imtis buvo suformuojama iš „Subalansuoto dabartinės latvių kalbos tekstyno“ (lat. *Līdzsvarotā mūsdienu latviešu valodas tekstu korpusa*), taigi latvių kalbos anotuotas tekstynas yra sudarytas iš atskirų pastraipų (o ne pilnų tekstų), be to, tekstynas suformuotas taip, kad į jį pakliūtų 2000 dažniausiai bendrajame tekstyne vartojamų veiksmažodžių (Grūzītis et al., 2018). Pastraipų atrankos įrankyje pastraipos parenkamos iš įvairių stilių tekstų, laikantis tokių proporcijų (jos šiek tiek skiriasi nuo „Subalansuoto dabartinės latvių kalbos tekstyno“ proporcijų): 60 proc. periodika, 20 proc. grožinė literatūra, 7 proc. moksliniai tekstai, 6 proc. teisiniai tekstai, 5 proc. *Saeimos* stenogramos ir 2 proc. kiti tekstai (Skadiņa et al., 2022).

Iš 1 lentelės matyti, kad anotuotas slovėnų rašytinės kalbos tekstynas *UD Slovenian SSJ*, pateiktas prie UD standarto tekstynų (plačiau Dobrovoljc et al., 2017) susideda iš keturių tipų tekstų (dydis 13 000 sakinių, 267 097 TV): grožinė literatūra, negrožinė literatūra, *Vikipedijos* tekstai, naujienų tekstai (kaip matyti, teisinių tekstų čia neįtraukta, o sakytinei kalbai yra atskiras UD standarto tekstynas). Tekstyno dalių proporcijos yra tokios: 80 proc. negrožinė literatūra (čia įeina profesionalams skirti moksliniai tekstai ir naujienos, kurių tikslinė grupė neprofesionalai); 9 proc. grožinės (drama, poezija, proza); 12 proc. *Vikipedijos* tekstų. *UD Slovenian SSJ* tekstynas sudarytas imant tekstus iš dviejų šaltinių: a) tam tikro dydžio imtys atrinktos iš bendrojo rašytinės slovėnų kalbos tekstyno *GIGAFIDA* – imtį sudarė pastraipos, kurios, kaip ir latvių kalbos tekstyne, suformuotos iš įvairių tekstų taip, kad būtų užtikrinta žanrų, autorių įvairovė (grožinė, negrožinė literatūra, periodika). Tekstyną sudaro 618 dokumentų, viename dokumente vidutiniškai apie 432 TV, 8 paragrafai ir apie 22 sakinius.

Taigi iš esmės numatyta naujų lietuvių kalbos gramatiškai anotuotų tekstynų sandara yra panaši į aptartus latvių ir slovėnų kalbų tekstynus, tačiau mūsų pasirinktas principas imti visus tekstus iš esmės skiriasi nuo latvių ir slovėnų taikyto imčių principo atrenkant teksto fragmentus. Imant ištraukas, galima palengvinti tekstyno sudarymo procesą (pvz., tekstinė medžiaga atrenkama ir paruošiama tekstynui automatizuotai, vadinasi, daug greičiau), o apsibrėžto dydžio imčių atranka palengvina balansavimą. Kita vertus, jeigu siekiama sudaryti gerokai didesnį anotuotą tekstyną, tada pilni tekstai duoda daugiau medžiagos, be to, iš pilnų tekstų sudaryto tekstyno pritaikymo galimybės yra platesnės, nes toks tekstynas gali būti puikus resursas pereinant prie semantinio lygmens (diskurso) analizės.

Tekstyno vieneto samprata

Automatinėje kalbos analizėje teksto apdorojimo etapas, einantis prieš anotavimą, yra skirtas tekstynui suskaidyti į tekstyno vienetus (angl. *tokenization*). Tekstyno vienetu (toliau – TV) laikomas žodis, tačiau gali būti ir kiti teksto elementai. Kinų, japonų kalbose, kur žodžių ribos nėra atskiriamos tarpais, teksto skaidymo problema didesnė, o pats procesas vadinamas segmentavimu (*The Handbook of Computational Linguistics and NLP*, 2010, p. 534). Vis dėlto ir tokiose kalbose, kaip lietuvių, anglų ar vokiečių, skaidant tekstyną į TV, neišvengiama probleminių atvejų. Šiame poskyryje aptarsime TV automatinio nustatymo problemas, paaiškinsime, kaip mūsų projekte yra suprantamas TV, kaip nustatomos TV ribos abejojant, kiek – vieną ar du – TV reikia skirti.

Anotavimo procesas susideda iš lemavimo (naudojami lemuokliai, morfologiniai anotatoriai) ir anotavimo (naudojami morfologiniai arba (ir) sintaksiniai anotatoriai)[33]. Morfologiniai anotatoriai automatiškai nustato analizuojamo žodžio (žodžio formos) gramatines charakteristikas. Tačiau morfologiškai anotuojant tekstą, anotavimo pažymos pridedamos ne tik prie žodžių, žodžių formų, bet ir kitų TV – simbolių, skyrybos ženklų. Žodis kaip tekstyno analizės vienetas kompiuterinėje lingvistikoje nėra lengva sąvoka (žodžio samprata lingvistikoje taip pat gana probleminė, žr. Haspelmath (2023)): skaidant tekstyną į TV susiduriama su atvejais, kai reikšminis vienetas susideda iš kelių žodžių (pvz., samplaikos *taip pat, kai kur*) arba vienas žodis apima du vienetus (vok. prielinksnis *zum*). Be to, skaidant tekstą į TV, susiduriama ne tik su žodžiais, bet ir su įvairiais kitokiais teksto elementais: simboliais, skaitmenimis, trumpiniais, skyrybos ženklais ir pan. Taigi sąvokos *tekstyno vienetas* ir *žodis* nėra sinonimiškos, žodžiai sudaro tik dalį tekstyno vienetų, todėl galima sakyti, kad TV yra dviejų tipų: žodžiai ir nežodžiai (angl. *non-words*).

Pasak Kübler ir Zinsmeister (2015, p. 7), TV riba galima laikyti nežodinius teksto elementus, nors tai ir nėra universali taisyklė. Pavyzdžiui, anglų kalbos žodis „don't“ dažnai skaidomas į du TV „do“ ir „not“; daugiažodis junginys kaip „in spite of“ gali būti suprantamas kaip sudarytas iš trijų TV, atskirtų tarpais (nors junginys yra semantiškai nedalomas); skyrybos ženklas gali būti atskirtas nuo žodžio, kuriam yra priskiriamas ir traktuojamas kaip atskiras TV (pavyzdžiui, filmo pavadinimą „Men in Black“ sudaro penki TV). Kita vertus, trumpiniuose skyrybos ženklas gali būti neatskiriamas nuo žodžio kaip atskiras TV – todėl „Mr.“ suprantamas kaip vienas TV, o ne du atskiri TV (Kübler ir Zinsmeister, 2015, p. 47).

Tekstyno skaidymas TV yra aktualus klausimas kalbos analizės įrankių kūrėjams (plg. Anthony (2013), Brezina (2018)). Pasak Anthony (2013, p. 149), viena problemų, su kuria susiduria tekstynų lingvistai, – kai pakartotinai atliekant tyrimą tie patys duomenys analizuojami su kita programine įranga ir nesutampa tyrimo rezultatai. Tokią situaciją lemia įvairios priežastys, viena pagrindinių – skirtingas TV traktavimas (ibid.). Pavyzdžiui, įkėlus originalios ir vertimų lietuvių kalbos tekstyną ORVELIT[34] į *AntConc*[35], rodoma, kad tekstyną sudaro 3 998 484 TV, o *LancsBox* v.6.x[36] nurodomi 4 019 354 TV. Anthony (2013, p. 149) teigia, kad tokie skirtumai priklauso nuo to, kokią TV sampratą taiko programos kūrėjas, pavyzdžiui, vienoje programoje TV bus laikomi tik raidžių nuo A iki Z deriniai, kitoje – dar ir skaitmenimis parašyti skaičiai; vienur skyrybos ženklai bus priskiriami TV, kitur ne. Kartais kūrėjai net leidžia pačiam vartotojui pasirinkti, ką laikyti TV. Pasak Brezina (2018, p. 39), aprašant konkretų tekstyną ir jo dydį, reikėtų paminėti, kokiu įrankiu naudotasi, kaip buvo skaičiuojami žodžių dažniai, kokia TV samprata vadovautasi (plg. tekstyno vienetą sudaro „kiekviena raidžių ar skaičių seka, skiriama skyrybos ar tarpo ženklo“).

Plačiai naudojamoje tekstynų analizės įrangoje *SketchEngine*[37] TV, kurio sampratą perimame ir mes, TV laikoma: žodžio forma (*going, trees, Mary, twenty-five*), skyrybos ženklai; skaičiai; trumpiniai (*3M, i600, XP, FB*), bet

[33] Anotuojant morfologiškai ir sintaksiškai, paprastai vyksta ir vienareikšminimo procesas, t. y. parenkama tame kontekste tikėtiniausia forma (tam naudojami statistiniai ar kitokie metodai).

[34] Vaičenonienė, J., Kovalevskaitė, J., ir Boizou, L. (2020). ORVELIT v3, CLARIN-LT digital library in the Republic of Lithuania, http://hdl.handle.net/20.500.11821/40.

[35] Anthony, L. (2024). *AntConc (Version 4.3.1) [Computer Software].* Tokyo, Japan: Waseda University. Žr. https://www.laurenceanthony.net/software.

[36] Brezina, V., Weill-Tessier, P., & McEnery, A. (2021). #LancsBox v. 6.x. [software package].

[37] Žr. https://www.sketchengine.eu/.

kurie kiti su tarpais rašomi teksto elementai. TV yra dviejų tipų: žodžiai ir nežodžiai. Žodžiai yra TV, kurie prasideda bet kokia abėcėlės raide (*book, working, Mary, T-shirt, post-1945, mp3, CO2*). Nežodžiams priskiriami tie TV, kurie prasideda ne raide: *25-hour*, *16-year-old*, *!important*, *3D*.

Rengiant naujus lietuvių kalbos anotuotus tekstynus, nuspręsta skyrybos ženklus laikyti TV, nes sintaksiškai anotuotuose tekstynuose skyrybos ženklai yra traktuojami kaip atskiri vienetai, atliekantys tam tikras sintaksines funkcijas. Pavyzdžiui, viename iš standartų, kuris naudojamas anotuojant šiame straipsnyje aprašomus tekstynus, – MULTEXT-East, – skyrybos ženklai turi atskiras pažymas, pvz., brūkšnys ir brūkšnelis žymimi „Th“, taškas – „Tp“, skliaustai – „Tr“, kablelis – „Tc“ ir pan. Tokios pačios sampratos, t. y. skyrybos ženklus laikyti TV, laikomasi ir kroatų, latvių, čekų, lenkų, ukrainiečių sintaksiškai anotuotuose tekstynuose[38].

Kai apsibrėžiama, kas laikoma TV, dar reikia nutarti dėl TV skyrimo, kai kyla abejonių, kiek TV skirti. Toliau aptarsime dažnai pasitaikančius, probleminius atvejus, kurie aktualūs lietuvių kalbai, pateikdamos paaiškinimų, kokios strategijos laikomės ir kaip tie patys atvejai traktuojami analizuojant kitas kalbas[39].

Vienu tekstyno vienetu laikoma:

1. vieną leksinį gramatinį vienetą sudarančios samplaikos, pvz., *bet kuris, kur nors, kažin kas, nuo mažens*.
2. Trumpiniai su tašku, pvz., *dr*. (daktaras), *m*. (metai), *a*. (amžius). Čekų, slovakų, ukrainiečių tekstynuose laikomi dviem TV, t. y. trumpinys ir taškas, o baltarusių, rusų, estų tekstynuose taškas laikomas atskiru TV tada, kai toks sutrumpinimas atsiranda sakinio pabaigoje.
3. Tarpu ar skyrybos ženklais atskirtos susijusios skaičių grupės, pvz., 20 000, 20.000, 2,5. Beveik visuose UD portale aprašytuose kitų kalbų (išskyrus makedonų) tekstynuose tokie atvejai laikomi vienu TV, o estų ir suomių tekstynuose tokios skaičių grupės laikomos kelianariais TV (angl. *multitoken word*).
4. Telefonų numeriai, trumpuoju būdu parašyta data, laikas (pvz., 8 111 22222, +370611122222; 2024 12 30, 2024-12-30, 2024/12/30; 10.30, 10:30; 10:30:05). Tokie atvejai laikomi vienu TV danų, islandų, švedų, lenkų, slovėnų, estų, italų, rusų tekstynuose, bet baltarusių tekstyne – atskiri TV.
5. Trupmenos, dydžių santykis, pvz., ½, 3/10. Vienu TV laikomi danų, islandų, švedų, armėnų tekstynuose.
6. Interneto svetainių nuorodos, pvz., https://universaldependencies.org/, https://www.lrt.lt/. Vienu TV laikoma ir olandų, baltarusių, rusų, norvegų tekstynuose.
7. Trumpiniai, sudaryti iš dviejų (ar daugiau) vienetų, rašomi be tarpų arba su tarpais, pvz., *t. y., t.y., i.e., et al*. Latvių, anglų, italų, portugalų, islandų, norvegų, švedų, airių, graikų kalbų tekstynuose tai irgi vienas TV. Čekų, slovakų, ukrainiečių, armėnų tekstynuose – atskiri TV. Tai reiškia, kad, tarkim, lietuviškas sutrumpinimas *t. y.* tų kalbų tekstynuose būtų išskaidytas į keturis TV: 1) t, 2) ., 3) y, 4) .
8. Apostrofu atskirtos asmenvardžių dalys, pvz., *O'Connor, McDonald's, John'as*. Taip pat traktuojama ir lenkų kalbos tekstyne.
9. Numeravimo atvejai, pvz.: 1., 1), 2.1, a), (b), iii), X. Latvių, islandų, estų, suomių, anglų kalbų tekstyne – taip pat vienas TV.
10. Iš skaitmenų, skyrybos ženklų ar simbolių sudaryti jausmaženkliai, pvz., :), ^_^ . Vienu TV laikoma baltarusių, estų tekstynuose, o suomių tekstyne laikoma kelianariais TV.
11. Žodžio dalys: galūnės, priešdėliai, priesagos, šaknys ir pan., pvz.: *susituokęs (-usi)* (vienu TV laikoma *-usi*), *mokinys, -ė* (vienu TV laikoma *-ė*), *-uoti, -aitė, -imas, ne-, be-, -im-*.

[38] Dėkojame Vytauto Didžiojo universiteto Lietuvių filologijos ir leidybos bakalauro programos studentei Sabinai Baltrūnaitei, surinkusiai informaciją apie tai, kas https://universaldependencies.org/ portale pateiktuose įvairių kalbų tekstynuose laikoma TV.

[39] Paminėtina, kad prie aptariamų TV aprašymo atvejų nurodomi kitų kalbų tekstynai tada, kai aptariama situacija buvo aiškiai aprašyta UD portale pateiktame TV apraše. Tikėtina, kad ir kitų nepaminėtų kalbų tekstynuose čia aprašyti atvejai laikomi vienu TV, tik jie nepateikti tų kalbų TV skyrimo strategijos aprašuose.

Keliais tekstyno vienetais laikoma:

1. Ribas nurodančios skaičių, sutrumpinimų, žodžių ir skyrybos ženklų kombinacijos, pvz., 1-10, 1–10, XX–XXI, Vilnius-Kaunas, Vilnius–Kaunas. Visais šiais atvejais yra trys TV, pavyzdžiui, 1) 1, 2) -, 3) 10; 1) Vilnius, 2) –, 3) Kaunas. Italų, armėnų, baltarusių tekstynuose tokie atvejai laikomi vienu TV.
2. Skaičių ir simbolių junginiai, pvz.: 2%, 2 %, 2$, 2 $.
3. Pasviruoju brūkšniu atskirti trumpiniai, savarankiški žodžiai, rodantys santykį arba alternatyvą, pvz., *km/val., atvykimo/išvykimo, ir / ar*.
4. Kitų kalbų dvigubi asmenvardžiai, pavadinimai, pvz., *Rolls-Royce, Jean-Jacque*. Lenkų ir baltarusių tekstynuose tokie atvejai laikomi atskirais TV. Latvių, olandų tekstynuose dvigubos pavardės laikomos vienu TV.
5. Lietuviški brūkšneliu sujungti asmenvardžiai, pvz., *Krėvė-Mickevičius, Antaitytė-Petraitienė, Jonas-Antanas*[40]; taip pat ir kiti bendriniai žodžiai, pvz., *lopšelis-darželis*.
6. Tikriniai pavadinimai (vietovardžiai, vardai, pavardės), į kuriuos įeina nekaitomi žodžiai, pvz.: *Rio de Žaneiras, Los Andželas, Niu Hamšyras, Van Bethovenas, Fon Bismarkas*. Estų tekstyne laikoma kelianariais TV.
7. Brūkšneliu trumpinami žodžiai, pvz., *e-paštas, e-sveikata*. Latvių, anglų, kroatų, serbų tekstynuose tokie atvejai laikomi atskirais TV.

Naudojant šią TV klasifikavimo sistemą, žodiniai ir nežodiniai vienetai pasiskirsto maždaug taip: apie 78 proc. sudaro žodiniai TV (be skyrybos ženklų, skaitmenų), o apie 22 proc. sudaro nežodiniai TV (skyrybos ženklai, skaitmenys, simboliai). Šie duomenys gauti iš dalies (apie 444 tūkst. TV) rengiamų naujų gramatiškai anotuotų tekstynų. Ši tekstynų dalis apima administracinius ir mokslinius tekstus. Vis dėlto panaši žodinių ir nežodinių TV proporcija turėtų būti ir kitų stilių tekstuose.

Remiantis atlikta analize ir priimtais sprendimais dėl TV, planuojama prie naujai rengiamų anotuotų tekstynų pateikti ir tekstyno skaidymo TV aprašą su TV skaidymo principais. Vadovaujantis gerąja praktika, pateikiant informaciją apie tekstynų sudarymą, svarbu paaiškinti ir taikytą TV sampratą, o tekstyno skaidymo TV strategija, kaip ir morfologinio anotavimo standartas, lemavimo principai ir pan., turi būti bazinė kiekvienos kalbos automatinės analizės įrankyno, vadinamojo NLP-PIPE, dalis (plg. latvių kalbos įrankyną (Znotiņs ir Cīrule, 2018) ar dokumentaciją prie anotuoto tekstyno *Latvian Treebank* v2.15.

Apibendrinimai ir rekomendacijos

Iki šiol parengti lietuvių kalbos gramatiškai anotuotų tekstynų ištekliai (morfologiškai anotuotas tekstynas MATAS, sintaksiškai anotuotas tekstynas ALKSNIS) yra nedideli ir nepakankami atsižvelgiant į lietuvių kalbos kompiuterizavimo ateities poreikius. Todėl ES *NextGenerationEU* projekte „Morfologiškai ir sintaksiškai anotuotų tekstynų modeliai apmokymui (auksiniai standartai)“ rengiami du nauji 10 mln. žodžių anotuoti tekstynai pagal tarptautinį UD standartą. Palyginus su kitų kalbų tekstynais, parengtais taikant UD standartą, tokios apimties tekstynai bus svarbūs resursai, sudarysiantys galimybes tobulinti lietuvių kalbos automatinę analizę ir jos įrankius.

Nauji tekstynai reprezentuos rašytinės lietuvių kalbos atmainą, todėl numatyta, kad juose, kaip ir „Dabartinės lietuvių kalbos tekstyne“, anotuotuose tekstynuose MATAS ir ALKSNIS, bus keturios dalys su tekstais iš grožinės, negrožinės (mokslinės), administracinės literatūros ir straipsniai iš internetinės periodikos. Planuojamos proporcijos labai priklausys nuo galimybių gauti grožinės, negrožinės literatūros tekstų autorių leidimus naudoti tekstus šiems tekstynams parengti, tačiau numatyta, kad, kaip įprasta ir kitų kalbų tekstynuose, apie 50 proc. sudarys publicistika, po 20 proc. – administraciniai ir moksliniai tekstai, o grožinės literatūros tekstai – apie 10 proc.

Rengiant straipsnyje minimus naujus tekstynus, siekiama patikslinti ir geriau dokumentuoti svarbius tekstyno rengimo etapus – vienas jų yra tekstyno skaidymo tekstyno vienetais (TV) etapas (angl. *tokenization*). *Tekstyno*

[40] Atkreipiame dėmesį, kad šis (ir keletas kitų) pavyzdžių parašytas nesilaikant bendrinės kalbos normų, bet tekstynuose tokie atvejai neredaguojami.

vienetas yra terminas, kurį šiame straipsnyje pasiūlėme vietoj angliško termino *token*. Nors tiek lietuvių, tiek kai kurių kitų kalbų tekstynų dydis gali būti nurodomas žodžiais, vis dėlto dažniausiai tekstynų dydis matuojamas TV, nes natūraliosios kalbos analizėje vienetai yra ne tik žodžiai, bet ir kiti elementai. Tekstyno vienetu laikome: žodžius (žodžių formas) ir nežodžius (skyrybos ženklus, skaitmenis, trumpinius, simbolius). Atskirų susitarimų reikia sprendžiant, kaip tam tikri teksto elementai turėtų būti vertinami – kaip vienas ar kaip du TV.

Kadangi TV samprata gali skirtis priklausomai nuo kalbos, tyrėjų taikytos strategijos, svarbu, kad konkrečios kalbos automatinės analizės įrankyne būtų aprašyta, kas laikoma žodžiu, kas laikoma TV, kokios TV skaidymo strategijos buvo laikomasi, kaip traktuojami nevienareikšmiai TV atvejai (brūkšneliu sujungti asmenvardžiai, trumpiniai su tašku ir kt.), kokie sprendimai dėl jų yra priimti. Naujų rengiamų tekstynų TV bus nustatyti pagal čia aprašytą TV sampratą. Tačiau kad ir kokia TV skaidymo strategija būtų taikoma konkrečiame tekstyne, rekomenduotume tekstynų dydžius nurodyti ir TV, ir žodžiais, kad tyrėjams būtų aiškus kalbinių duomenų kiekis.

Interesų konfliktas

Autoriai deklaruoja, kad interesų konflikto dėl šio straipsnio publikavimo nėra.

Literatūra

1 Anthony, L. (2013). A critical look at software tools in corpus linguistics. Linguistic Research, 30, 141–161. https://doi.org/10.17250/khisli.30.2.201308.001

2 Bielinskienė, A., Boizou, L., Kovalevskaitė, J., & Rimkutė, E. (2016). Lithuanian Dependency Treebank ALKSNIS. Proceedings of the Seventh International Conference Baltic HLT 2016, Amsterdam: IOS Press, 107–114. Retrieved January 10, 2025, http://ebooks.iospress.nl/volumearticle/45523

3 Bielinskienė, A., Boizou, L., & Rimkutė, E. (2017). Lietuvių kalbos morfologiškai ir sintaksiškai anotuoti tekstynai. Bendrinė kalba, 90, 1–30. Retrieved January 10, 2025, http://www.bendrinekalba.lt/Straipsniai/90/Bielinskiene ir kt_BK_90_straipsnis.pdf.

4 Boizou, L., & Zamblera, F. (2014). Syntactic engine for the Lithuanian language. Proceedings of the 6th international conference, Baltic HLT 2014, Amsterdam: IOS Press, 69–74. Retrieved January 10, 2025, http://ebooks.iospress.nl/publication/38006.

5 Boizou, L., Kapočiūtė-Dzikienė, J., & Rimkutė, E. (2018). Deeper error analysis of Lithuanian morphological analyzers. Human language technologies - the Baltic perspective: proceedings ofthe 8th international conference Baltic HLT, Amsterdam: IOS Press, 18–25.

6 Brezina, V. (2018). Statistics in Corpus Linguistics: A practical guide. Cambridge: University Press. https://doi.org/10.1017/9781316410899

7 Dadurkevičius, V. (2017). Lietuvių kalbos morfologija atvirojo kodo „Hunspell“ platformoje. Bendrinė kalba, 90, 1–15. Retrieved January 10, 2025, from https://etalpykla.lituanistika.lt/fedora/objects/LT-LDB-0001:J.04~2017~1525788525052/datastreams/DS.002.0.01.ARTIC/content

8 de Marneffe, M.-C., Manning, Ch. D., Nivre, J., & Zeman, D. (2021). Universal Dependencies. Computational Linguistics, 47(2), 255–308. doi: https://doi.org/10.1162/coli_a_00402

9 Dobrovoljc, K. (2022). Spoken language treebanks in Universal Dependencies: An overview. Proceedings of the Thirteenth Language Resources and Evaluation Conference, Marseille, France, 1798–1806. Retrieved January 10, 2025, from https://aclanthology.org/2022.lrec-1.191/

10 Dobrovoljc, K., Erjavec, T., & Krek, S. (2017). The Universal Dependencies Treebank for Slovenian. Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing, Valencia, Spain, 33–38. Retrieved January 10, 2025, from https://doi.org/10.18653/v1/W17-1406

11 Hajič, J., Hajičová, E., Mikulová, M., & Mírovský, J. (2017). Prague Dependency Treebank. Handbook of Linguistic Annotation. Ide, N., Pustejovsky, J. (Eds.), Springer, 555–594. Retrieved January 10, 2025, from https://doi.org/10.1007/978-94-024-0881-2_21

12 Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M. Štěpánek, J., & Štěpánková, B. (2020). Prague Dependency Treebank - Consolidated 1.0. Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 5208–5218.

Retrieved January 10, 2025, from https://aclanthology.org/2020.lrec-1.641/

13 Clark, A., Fox, Ch., & Lappin, Sh. (Eds.). (2010). The Handbook of Computational Linguistics and NLP. Wiley-Blackwell. https://doi.org/10.1002/9781444324044

14 Haspelmath, M. (2023). Defining the word. WORD, 69(3), 283–297. https://doi.org/10.1080/00437956.2023.2237272

15 Grumadienė, L. (2002). Dabartinės rašomosios lietuvių kalbos dažninis žodynas ir jo bazė. Acta linguistica Lithuanica, 46, 19–37. Retrieved January 10, 2025, from https://etalpykla.lituanistika.lt/object/LT-LDB-0001:J.04~2002~1367164490423/J.04~2002~1367164490423.pdf

16 Kapočiūtė-Dzikienė, J., Rimkutė, E., & Boizou, L. (2017). A Comparison of Lithuanian morphological analyzers. Proceedings of the TSD 2017: Text, Speech, and Dialogue: 20th International Conference. Ekštein, K., Matoušek, V. (Eds.), Springer, 47–56. https://doi.org/10.1007/978-3-319-64206-2_6

17 Kübler, S., & Zinsmeister, H. (2015). Corpus Linguistics and Linguistically Annotated Corpora. London: Bloomsbury Publishing.

18 Nivre, J. (2015). Towards a universal grammar for natural language processing. Computational Linguistics and Intelligent Text Processing, vol. 9041. Gelbukh, A. (Ed.), Springer International Publishing, 3–16. https://doi.org/10.1007/978-3-319-18111-0_1

19 Pretkalniņa, L., Rituma, L., & Saulīte, B. (2018). Deriving Enhanced Universal Dependencies from a Hybrid Dependency-Constituency Treebank. Text, Speech, and Dialogue, vol. 11107, 95–105. https://doi.org/10.1007/978-3-030-00794-2_10

20 Rimkutė, E. (2024). Ką gali Morfuoklis? Gimtoji kalba, 7, 8–15.

21 Rimkutė, E., & Daudaravičius, V. (2007). Morfologinis dabartinės lietuvių kalbos tekstyno anotavimas. Studies about Languages / Kalbų studijos, 11, 30–35. https://www.lituanistika.lt/content/17493

22 Rimkutė, E. (2006). Morfologinio daugiareikšmiškumo ribojimas kompiuteriniame tekstyne. Daktaro disertacija. Kaunas: Vytauto Didžiojo universitetas.

23 Rituma, L., Pretkalniņa, L., Saulīte, B., Nešpore-Bērzkalne, G., Grūzītis, N., & Znotiņš, A. (2024). LVTB - Latvian Treebank v2.15 (2024-11-15), CLARIN-LV digital library at IMCS, University of Latvia. Retrieved January 10, 2025, from http://hdl.handle.net/20.500.12574/112

24 Rituma, L., Saulīte, B., & Nešpore-Bērzkalne, G. (2019). Latviešu valodas sintaktiski marķētā korpusa gramatikas modelis. Language: Meaning and Form, 10, 200–216. https://doi.org/10.22364/vnf.10.16

25 Rosén, V., De Smedt, K., Meurer, P. & Dyvik, H. (2012). An open infrastructure for advanced treebanking. META-RESEARCH Workshop on Advanced Treebanking at LREC2012, Hajič, J., De Smedt, K., Tadić, M., Branco, A. (Eds.), Istanbul, Turkey, 22–29. https://doi.org/10.1007/s10579-016-9356-5

26 Skadiņa, I., Saulīte, B., Auziņa, I., Grūzītis, N., Vasiļjevs, A., Skadiņš, R., & Pinnis, M. (2022). Latvian Language in the Digital Age: The Main Achievements in the Last Decade. Baltic Journal of Modern Computing, 10(3), 490–503. https://doi.org/10.22364/bjmc.2022.10.3.21

27 Tour de CLARIN 2024: Interview with Erika Rimkutė and Virginijus Dadurkevičius. Retrieved January 10, 2025, from https://www.clarin.eu/blog/tour-de-clarin-interview-erika-rimkute-and-virginijus-dadurkevicius.

28 Universal Dependencies. Retrieved January 10, 2025, from https://universaldependencies.org/

29 Utka, A. (2009). Dažninis rašytinės lietuvių kalbos žodynas: 1 milijono žodžių morfologiškai anotuoto tekstyno pagrindu. Kaunas: Vytauto Didžiojo universiteto leidykla. Retrieved January 10, 2025, https://www.vdu.lt/cris/entities/publication/91520333-855f-4ba9-a094-968a0366aaa9

30 Zeman, D. (2018). The World of Tokens, Tags and Trees. Studies in Computational and Theoretical Linguistics. Retrieved January 10, 2025, from https://ufal.mff.cuni.cz/books/preview/2018-zeman_full.pdf

31 Zinkevičius, V. (2000). Lemuoklis – morfologinei analizei. Darbai ir dienos, 24, 245–274. Retrieved January 10, 2025, from http://etalpykla.lituanistikadb.lt/fedora/objects/LT-LDB-0001:J.04~2000~1367182714637/datastreams/DS.002.0.01.ARTIC/content https://doi.org/10.7220/2335-8769.24.15

32 Znotiņš, A., & Cīrule, E. (2018). NLP-PIPE: Latvian NLP Tool Pipeline. Human Language Technologies – The Baltic Perspective, Amsterdam: IOS Press, 183–189.

Abstract

Jolanta Kovalevskaitė, Erika Rimkutė, Jurgita Vaičenonienė
Developing new annotated corpora for Lithuanian: Compilation issues

The currently available Lithuanian grammatically annotated corpora (the morphologically annotated corpus MATAS and the syntactically annotated corpus ALKSNIS) are insufficient to meet the growing demands of Lithuanian language processing. Therefore, within the framework of the European Union's NextGenerationEU project, "Morphologically and syntactically annotated text models for training (gold standards)", two new corpora are being developed in accordance with the international Universal Dependencies (UD) standard. These new corpora (10 million tokens each) will represent the written variant of the Lithuanian language and, similar to the Corpus of Contemporary Lithuanian Language and other annotated corpora, they will consist of four sub-corpora containing texts from fiction, non-fiction (scientific literature), administrative texts, and online journalistic articles.

Following the presentation of the annotated corpora MATAS and ALKSNIS, as well as the UD corpora of other languages, this paper discusses various aspects of the annotated corpora under development. First, the corpora structure, balance and sampling are described. Second, tokenization, essential for the automatic corpus analysis is examined, and third, the concept of *token* is discussed. Although corpora size can be expressed in words, it is most commonly measured in tokens (suggested Lithuanian translation: *tekstyno vienetas*), as corpora include not only words, but also non-words, e.g., punctuation marks, numerals, abbreviations, symbols, etc. Since the elements included under the concept of token may vary depending on the language and researchers' decisions, the article discusses what is considered a token for Lithuanian within the boundaries of the current project.

About the Authors

JOLANTA KOVALEVSKAITĖ

Dr., vyresnioji mokslo darbuotoja, Skaitmeninių išteklių ir tarpdisciplininių tyrimų institutas, Vytauto Didžiojo universitetas, Lietuva

Mokslinių interesų sritys
Tekstynų lingvistika, pastovieji žodžių junginiai, leksikografija

Adresas
V. Putvinskio g. 23-216, Kaunas, LT-44243, Lietuva

El. paštas
jolanta.kovalevskaite@vdu.lt

Orcid
0000-0001-7449-5518

ERIKA RIMKUTĖ

Doc. dr., vyresnioji mokslo darbuotoja, Skaitmeninių išteklių ir tarpdisciplininių tyrimų institutas, Vytauto Didžiojo universitetas, Lietuva

Mokslinių interesų sritys
Tekstynų lingvistika, kompiuterinė lingvistika, pastovieji žodžių junginiai, automatinė gramatinė analizė

Adresas
V. Putvinskio g. 23-216, Kaunas, LT-44243, Lietuva

El. paštas
erika.rimkute@vdu.lt

Orcid
0000-0003-0858-8593

JURGITA VAIČENONIENĖ

Doc. dr., mokslo darbuotoja, Skaitmeninių išteklių ir tarpdisciplininių tyrimų institutas, Vytauto Didžiojo universitetas, Lietuva

Mokslinių interesų sritys
Tekstynų lingvistika, vertimo studijos, mokslinių tyrimų infrastruktūros

Adresas
S. Daukanto g. 27-204, Kaunas, LT-44249, Lietuva

El. paštas
jurgita.vaicenoniene@vdu.lt

Orcid
0000-0002-4440-896X