ktu faculty of social sciences, arts and humanities 1922

SAL 39/2021
Research Journal
Studies about Languages
pp. 93-110
ISSN 1648-2824 (print)
ISSN 2029-7203 (online)
DOI 10.5755/j01.sal.1.39.29258

LINGUISTICS / KALBOTYRA

Sintaksinio sudėtingumo analizė anotuotame lietuvių kalbos tekstyne priklausomybių nuotolio metodu

Received 06/2021 | Accepted 10/2021

crossref http://dx.doi.org/10.5755/j01.sal.1.39.29258

HOW TO CITE: Ožeraitis, V. (2021). Sintaksinio sudėtingumo analizė anotuotame lietuvių kalbos tekstyne priklausomybių nuotolio metodu. *Studies about Languages / Kalbų studijos, 39*, 93–110. http://doi.org/10.5755/j01.sal.1.39.29258

Sintaksinio sudėtingumo analizė anotuotame lietuvių kalbos tekstyne priklausomybių nuotolio metodu

Analysis of Syntactic Complexity in the Annotated Lithuanian Language Corpus by the Method of Dependency Distance

VYTAUTAS OŽERAITIS, Vytauto Didžiojo universitetas, Lietuva

Santrauka

Sintaksinis sudėtingumas yra visoms kalboms būdinga ypatybė, labai bendrai apibūdinama kaip sakinio (ar teksto) ir jo elementų įmantrumą, detalumą, sąrangą bei jungimosi modelius sudėtingumo aspektu pateikiantis įvertis. Lietuvių kalboje sintaksinis sudėtingumas nėra plačiai analizuotas. Sintaksinio sudėtingumo tyrimus apsunkina nenusistovėjusi termino apibrėžtis ir skirtingų jo apskaičiavimo metodų gausa. Šiame straipsnyje pristatomas sintaksinio sudėtingumo tyrimas sintaksiškai anotuotame lietuvių kalbos tekstyne ALKSNIS, naudojant sintaksinės priklausomybės nuotolio metodą, paremtą *Dependency Locality* teorija. Straipsnyje aprašoma sintaksinio sudėtingumo samprata, pristatomi sintaksinio sudėtingumo tyrimų principai, jų aktualumas ir pritaikomumas, pristatomi ir aptariami sintaksinio sudėtingumo rezultatai tekstyne, analizuojami pasirinkto metodo privalumai ir trūkumai. Šiuo tyrimu siekiama papildyti lietuvių kalbos sintaksinio sudėtingumo analizės lauką.

Sintaksinio sudėtingumo analizei naudojami du rodmenys – vidutinis priklausomybės nuotolis ir modifikuotas vidutinis priklausomybės nuotolis, sintaksiškai anotuotame tekstyne apskaičiuojant sintaksinių sakinio priklausomybės ryšių distanciją. Tyrime analizuojami visi tekstyno sakiniai, nustatomas atskirų tekstų ir tekstyno dalių sintaksinis sudėtingumas. Detaliau analizuojant paskirus sakinius, išryškėja tiek šių metodų trūkumai sintaksinio sudėtingumo analizei, tiek jų priklausomybė nuo tikslios ir nuoseklios anotavimo schemos. Analizuojant duomenis išryškėja poreikis į sintaksinio sudėtingumo sampratą bei formulę įtraukti sakinių tarpusavio sąsajų svertus. Nustatyta, kad modifikuoto vidutinio priklausomybės nuotolio formulėje įtraukta sakinio viršūnės pozicija galimai iškreipia rezultatus, todėl šią formulę reikėtų toliau tikslinti. Tyrimo metu nustatytos sakinių ir tekstų sudėtingumo ribos laikomos tik orientacinėmis, tiksliau joms apibrėžti siūlomi papildomi kokybiniai tyrimai ir eksperimentai.

RAKTAŽODŽIAI: sintaksinis sudėtingumas, tekstynų lingvistika, sintaksinės priklausomybės nuotolis, lietuvių kalba, tekstynų anotavimas.

Įvadas

Sudėtingumo klausimas, nors nuolatos sprendžiamas, vis dar aktualus. Atrodo, kad sudėtingus ir paprastus dalykus galima skirti vien iš intuicijos, tačiau kai prireikia įvardyti ir apibrėžti jų bruožus, kyla įvairių problemų. Rescheris (1998) rašė:

> *Pasaulio sudėtingumas yra svarbus gyvenimo faktas, mums turintis esminių pasekmių. <...> Bet kuri sistema ar procesas – bet kas, ką galima laikyti iš tarpusavyje susijusių dalių sudaryta struktūra – turės vienokį ar kitokį sudėtingumo laipsnį. Daugiau ar mažiau, sudėtingumas yra paplitęs visoje realaus (ir fiktyvaus) pasaulio teritorijoje (p. 1).*

Šia ypatybe pasižymi ir kalba, komunikacija, kuriami ir suprantami tekstai. Greitame informacijos sraute nustatyti ir tirti sudėtingumą svarbu ir reikalinga. Kalbos ir jos produktų sudėtingumo tyrimai reikalingi vystant kalbines technologijas, mokantis ir mokant užsienio kalbų, kuriant efektyvų reklaminį ir komunikacijos turinį, tiriant kognityvinius gebėjimus, diskutuojant apie žanrų ypatybes ir ribas. Šios interesų sritys apima skirtingas mokslines disciplinas, kuriose kalbos sudėtingumas užima svarbią vietą: nuo diachroninės lingvistikos ir kalbų tipologijos, iki stilistikos, kalbos įgijimo ir vystymosi, kompiuterinės lingvistikos. Kokios pakraipos tyrimas bebūtų, svarbu apibrėžti sintaksinio sudėtingumo esmę ir pasirinkti jo analizės būdą, suprasti kas ir kokiu tikslu yra matuojama bei turėti tinkamus įrankius tai atlikti.

Mokslinėje literatūroje galima rasti daug ir įvairių sintaksinio sudėtingumo matavimo būdų, dažnai pasirenkamų pagal tyrimo tikslą ir pačią sąvokos sampratą, kuri vis dar nėra iki galo nusistovėjusi. Straipsnyje pristatomas vienas iš sintaksinio sudėtingumo matavimo būdų. Aprašomo tyrimo **tikslas** – pritaikyti sintaksinės priklausomybės nuotolį analizuojant sintaksinį sudėtingumą anotuotame lietuvių kalbos tekstyne ALKSNIS, įvertinti rezultatus, metodikos privalumus ir trūkumus. Šio tyrimo **objektas** – sintaksinis sudėtingumas anotuotame lietuvių kalbos tekstyne. Tyrimu siekiama papildyti lietuvių kalbos sintaksinio sudėtingumo analizės lauką, pristatyti inovatyvų, naują, nuoseklų ir kalbų bei tekstų tarpusavio lyginimą palengvinantį metodą. Verta pastebėti, kad vieną naujausių panašaus pobūdžio tyrimų atliko D. Kalinauskaitė (2018), tačiau šie du darbai skiriasi tiek sudėtingumo samprata, tiek jo lygmenų apibrėžimu, tiek naudota metodologija. Keliami **uždaviniai:** pristatyti *sintaksinio sudėtingumo* sampratą; apskaičiuoti kiekvieno ALKSNYJE esančio sakinio ir teksto sudėtingumo rodiklius; juos aprašyti, palyginti ir įvertinti metodikos privalumus ir trūkumus; nurodyti galimas tolesnio tokios krypties tyrimo gaires. Darbą sudaro dvi dalys. Pirmoje pateikiama glausta literatūros apžvalga, pristatomas sintaksinis sudėtingumas ir pagrindžiamas metodo pasirinkimas. Antroje dalyje aptariami tyrimo rezultatai.

Literatūros apžvalga

Sintaksinio sudėtingumo sąvoka lingvistų darbuose įsitvirtinusi, nors vietomis vis dar kelia tam tikrų abejonių. Pats sintaksinis sudėtingumas, jį analizuojančioje literatūroje tai akcentuojama gana retai, yra sudėtinė dalis kur kas platesnės apimties reiškinio, vadinamo lingvistiniu sudėtingumu (šiame darbe lingvistinis ir kalbos sudėtingumas suprantami kaip sinonimai). Net kalbant labai bendrai, sunku pateikti vieną aiškų, detalų ir nuoseklų lingvistinio sudėtingumo apibrėžimą. Žiūrint įprastai, jis galėtų apibrėžti kalbos kaip darinio ar sistemos sudėtingumą žodyninėmis reikšmėmis – *susidedantis iš daugelio dalių, elementų, sudėtinis, nevientisas* ar *komplikuotas, sunkiai suvokiamas* (LKŽ). Vis dėlto tyrėjų darbuose iškeliama ne viena problema, kuri apsunkina paprastą ir tiesmuką sąvokos apibrėžimą.

Steger ir Schneider (2012) pabrėžia, kad vieno, bendrai priimto ir nuo atskiros teorijos nepriklausomo lingvistinio sudėtingumo apibrėžimo nėra. Pastebima, kad beveik neįmanoma sutarti, ar apskritai gali būti absoliutus kalbos sudėtingumas, pamatuojamas vienu nuo konteksto nepriklausomu rodikliu, o galbūt sudėtingumas gali būti vertinamas tik lyginant vieną sistemą su kita. Negana to, išreiškiama abejonė, ar sudėtingumą apskritai galima vertinti holistiškai, užuot vertinus ir lyginus tik paskiras visumos dalis. Bene daugiausia abejonių išsakoma dėl pačių sudėtingumo matavimo būdų. Kaip matuoti ir aprašyti lingvistinį sudėtingumą? Kiekybiniais terminais, laikantis nuostatos „daugiau yra sudėtingiau“ ar kokybiškai, pasitelkiant, pvz., psicholingvistinę analizę?

Kaip jau minėta, lingvistinis sudėtingumas nėra dažnai minimas rašant apie viena jo dalių laikytiną sintaksinį sudėtingumą. Tai sukelia savų problemų, kadangi atsiranda klaidingos interpretacijos galimybė, sunku suprasti, kas kiekvienu skirtingu atveju traktuojama kaip *sudėtingumas*. Šią problemą išryškina Ma ir Wang (2019), paminėdami, kad lingvistinio sudėtingumo tyrimams trūksta sisteminio požiūrio, bendro sutarimo dėl sudėtingumo sampratos. Taip pat abejojama dėl sudėtingumo, fiziškai apčiuopiamos ir pamatuojamos pasaulio sistemų savybės, formalizuotos išraiškos galimybių apimti ir perteikti kalbinio sudėtingumo visumą. Anot straipsnio autorių, pagrindiniais lingvistinio sudėtingumo požymiais galima laikyt nelijiniškumą ir organizacinę sistemą sudarančių elementų tvarką, pasireiškiančią lygmenų, dimensijų ir skirtingų būsenų gausa bei aukštu kardinalumu. Kitaip

tariant, sudėtingumas išauga kartu su minimaliu informacijos aprašymo kiekiu ir išaugusiais tos informacijos apdorojimo ištekliais/kaina. Pirmasis aspektas yra siejamas su informacijos teorija ir sudėtingumą pirmiausia apibūdina kiekybiniais rodikliais. Paprastai tariant, išsireiškimo (sakinio, pasakymo, žodžio) ilgis nurodo ir didesnį sudėtingumą, kadangi kiekvienas papildomas elementas kartu atsineša tiek atitinkamas semantines ypatybes, tiek potencialią sąveiką su kitais elementais, taip plečiant sintaksinių ryšių tinklą ir keliant jo sudėtingumą. Nors labiausiai paplitęs ir lingvistikos tyrimuose įsigalėjęs, vien kiekybinis sudėtingumo analizės pasirinkimas kelia pagrįstų abejonių. Kaip pažymima Biber et. al. (2020), Larsson ir Kaatari (2020) darbuose, sudėtingumo susiejimas vien su analizuojamo vieneto ilgio, kiekio ar gausos rodikliais nėra pakankamas, kadangi tokiu būdu neatsižvelgiama į skirtingų lingvistinių kategorijų ypatybes, arba kelis kartus matuojami tie patys dalykai. Pvz., vienodo ar panašaus ilgio sakiniai gali būti sudaryti iš kelių šalutinių sakinių arba kelių vienas į kitą įterptų išplėtotų žodžių junginių be šalutinio sakinio. Nors vienetų ilgis panašus, iš esmės skiriasi lingvistinio sudėtingumo rūšys. Informacijos apdorojimo išteklius/kainą galima sieti su vartotojo (skaitytojo, klausytojo) patiriamu sunkumu skaitant, girdint, suvokiant tekstą. Kalbant paprasčiau, šiuo būdu sudėtingumas vertinamas ne iš kalbos produkto (konkrečiu atveju – teksto) kūrėjo, bet iš jo vartotojo perspektyvos. Gibson (1998), pristatydamas priklausomybių pozicijos teoriją (ang. *Dependency Locality Theory*[1]) teigia, jog kalbos (teksto) suvokimo procesui svarbūs du reiškiniai (ang. *resources*) – saugykla ir integracija (ang. *storage* ir *integration*). Tai daugiausia susiję su žmogaus atminties galimybėmis ir apkrovomis. Anot teorijos autoriaus, sudėtingumas priklauso nuo informacijos (konkrečiu atveju prijungiamų žodžių ar sakinio dalių), kurią reikia išlaikyti atmintyje kiekio ir naujos informacijos apimčių. Tokiai analizei svarbūs sintaksiškai anotuotų sakinių priklausomybių medžiai ir kalbos vieneto vieta juose. Kuo didesnė vidutinė priklausomybės distancija, tuo daugiau išteklių reikalaujama sakiniui suprasti, tuo didesnis jo sudėtingumas.

Pristatyta teorija lingvistinį ir sintaksinį sudėtingumą pirmiausia sieja su pačiais kalbančiaisiais ar rašančiaisiais. Vis dėlto čia būtų pravartu bent trumpai paminėti žymesnius darbus, kurie į lingvistinį sudėtingumą žvelgia kaip į bendrą kalboms būdingą reiškinį, bando jį apibrėžti ar klasifikuoti. Lingvistinio sudėtingumo tyrimus galima skirstyti pagal 1998 m. Rescherio pateiktą klasifikaciją. Savo darbe autorius bendrai svarsto apie sudėtingumo sampratą ir jos įvairovę, skirdamas kelias skirtingas sudėtingumo kategorijas (ang. *modes of complexity*): epistemine, ontologinę ir funkcinę. Kiekviena iš jų skirstoma į dar smulkesnes kategorijas, kurias, pasak Karlsson (2014), atitinka pagrindinės lingvistinio sudėtingumo tyrimų kryptys.

Galima paminėti tris svarbius leidinius, kurių autorių darbuose sudėtingumas dažniausiai tiriamas kaip visos kalbos sistemos savybė. Šie darbai, pagal Rescherio klasifikaciją, galėtų būti laikomi kaip tiriantys *aprašomąjį, struktūrinį* ir *taksonominį* (ang. descriptive, constitutional, taxonomic) sudėtingumą, kuris pasižymi sistemos aprašo apimtimi, sistemą sudarančių elementų kiekiu ir jų įvairove. Kortman ir Szmrecssanyi (2012), Miestamo et. al. (2008) ir Sampson et. al. (2009) sudarytuose leidiniuose tiriamos kalbos ir jų atmainos (pvz., anglų ir kreolinių kalbų lyginimas), aptariami kalbų lingvistinio sudėtingumo teoriniai ir metodologiniai aspektai, tipologiniai ir kalbų kaitos klausimai.

Funkcinio sudėtingumo aspektu daugiausia analizuojama viena kuri sistemos dalis (konkrečiu atveju kalbama apie sintaksinį sudėtingumą). Į sudėtingumą žvelgiama iš kalbos vartotojo perspektyvos. Daugiausia dėmesio tam skiriama antrosios ar užsienio kalbos įgijimo ir mokymosi tyrimuose. Pallotti (2015) sudėtingumą siūlo vertinti dvejopai: kaip nepriklausomą kintamąjį, nurodantį, kiek komunikacijos užduotis yra paprasta ar sudėtinga (daugiausia siejama su *sunkumo* reikšme) ir priklausomą kintamąjį, apibūdinantį lingvistinio produkto (žodžio, sakinio, teksto) bruožus – komponentų kiekį, ilgį, ryšių kiekį ir ilgį. Tokią pat skirtį aptaria Bulte ir Housen (2012). Čia pateikiama ir apibendrinta kalbos sudėtingumo kategorizacija, kuri, nors visų pirma skirta antrosios kalbos įsisavinimui, tinkama ir funkciniam sudėtingumui tyrinėti kitais pasirinktais aspektais.

Iki šiol aptarta, kad sintaksinis sudėtingumas yra lingvistinio sudėtingumo dalis. Tai siauresnė plačios apimties tyrimų sritis, daugiausia dėmesio skirianti ne sisteminiam sudėtingumo tyrimui, bet kalbos vartotojo santykiui su kalba apibūdinti (iš supratimo, mokymosi, įsisavinimo ir kūrimo perspektyvos). Sintaksinį sudėtingumą galima laikyti analizuojamo teksto ypatybe (ang. feature), nurodančia jo sunkumo arba išmanumo (ang. *sophisticated*

[1] Vertimas – šio darbo autoriaus.

vert. – autoriaus) laipsnį. Išmanumu galima laikyti tiek kalbinių kompetencijų demonstravimą, tiek gebėjimą kurti tekstą atsižvelgiant į kalbinę ir komunikacinę situacijas. Nors tarp sunkumo ir išmanumo galima įžvelgti skirtį, tačiau tuo pačiu verta svarstyti apie galimai glaudų jų ryšį – kiek gebama kurti tokį tekstą, kuris adresatui nebūtų per daug sunkus ar sudėtingas. Kitaip tariant, sukurtas tekstas yra komunikacijos produktas, todėl jo sudėtingumas taip pat gali būti vertinamas ir iš priimančiojo perspektyvos.

Yra sukurta nemažai skirtingų būdų, kaip išmatuoti sintaksinį sudėtingumą. Daugiausia jų skirta antrosios kalbos įsisavinimo tyrimams, analizuojant sintaksinio sudėtingumo ir kalbinių kompetencijų sąsajas. Kalbant apie skirtingus sintaksinio sudėtingumo matavimo metodus, svarbu paminėti Lu (2011) tyrimą, kuriame apžvelgti ir įvertinti 14 įprastai tokio pobūdžio darbuose naudojamų matavimo metodų. Kaip nurodo autoriai, tai tik dalis iš beveik 100 iki tol užfiksuotų sudėtingumo matavimo metodų. Jie suskirstyti į penkias grupes, priklausomai nuo to, kas yra matuojama – sakinio ar pasakymo ilgis, šalutinio sakinio ilgis, santykinis šalutinių sakinių kiekis, veiksmažodinės frazės ar sudėtingų vardažodinių frazių kiekis sakinyje/sakinio dėmenyje. Visi šie metodai susiję su skirtingų tipų elementų ar ryšių tarp tų elementų kiekio nustatymu ir atitinka Bulte ir Housen (2012) suformuotą sudėtingumo sampratą. Dėl šių metodų, jų tikslumo ir naudos nuomonės skiriasi. Szmrecsanyi (2004) argumentuoja, kad pats paprasčiausias metodas yra ir pats efektyviausias (pirmiausia ekonomijos ir spartos požiūriu) – anot autoriaus, žodžių kiekio tekste matavimas leidžia teksto sintaksinį sudėtingumą įvertinti beveik taip pat tiksliai, kaip ir įmantresni metodai, ženkliai sutaupant laiko, kadangi teksto nereikia paruošti išsamiai kompiuterinei analizei. Kur kas gilesnis požiūris į sintaksinio sudėtingumo matavimo priemones ir metodus pateiktas Biber et al. (2020). Kaip jau minėta, čia iškeliama idėja, jog vien komponentų ilgiu paremti sudėtingumo metodai gali atskleisti tik dalį reiškinio visumos, kurios kitai daliai būtini analitiniai tyrimo būdai. Pagrindinius sintaksinio sudėtingumo metodų trūkumus ir apribojimus aptaria Bulte ir Housen (2012), 36 sudėtingumo tyrimus ir juose taikytus metodus išsamiai apžvelgia Jagaiah et al. (2020) ir Ortega (2015).

Kaip jau minėta, sintaksinio sudėtingumo matavimai dažniausiai taikomi antrosios ar užsienio kalbos mokymo ir įgijimo tyrimuose. Pasitelkus tekstynų lingvistikos metodus, lyginami gimtakalbių ir negimtakalbių kuriami akademiniai tekstai bendrojo lavinimo mokyklose ir studijų aukštosiose mokyklose metu, analizuojant kalbos įsisavinimo lygį, bandant įvertinti kuriamų tekstų kokybę, pademonstruojant iškylančius sunkumus ir ieškant efektyviausių mokymo metodų. (Ai, Lu 2013; Ortega 2003; Martinez 2017; Wu et al. 2015; Casal, Lee 2019; Ansarifar et al. 2018; Díez-Bedmar, Pérez-Paredes 2020; Yin et al., 2021; Jiang, Biber, Gray 2016, Jin et al. 2020). Lyginamas skirtingo stiliaus, žanro ir funkcijų tekstų sintaksinis sudėtingumas (Beers, Nagy 2011); tiriama, koks sintaksinio sudėtingumo ryšys egzistuoja tarp mokinių skaitomų (suvokimo, sunkumo prasme) ir priimamų tekstų (Barrot 2015); tiriamos sintaksinio sudėtingumo matavimo priemonių galimybės automatiškai nustatyti kuriamo teksto kokybiškumą (Shadloo et al., 2019; Thongyoi, Poonpon 2020); kaip stilistinę teksto ypatybę sintaksinį sudėtingumą aptaria Stauder ir Ustaszewski (2020), sintaksinis sudėtingumas analizuojamas kaip viena iš vertimo problemų (Borillo 2000), tirta jo svarba reklaminių tekstų poveikiui (Lowrey 1998; Liu, Afzaal 2021), nemažai dėmesio skiriama sintaksinio sudėtingumo ir natūralios kalbos apdorojimo programų sąveikai tirti, ieškant efektyviausio sudėtingumo vertinimo ir pritaikymo priemonių – kuriant santraukas, automatiškai trumpinant ir paprastinant tekstus, tobulinant dirbtinį intelektą (Evans, Orasan 2019; Chen, Zechner 2011).

Aukščiau paminėti tyrimai pademonstravo sintaksinio sudėtingumo taikymo galimybes. Naudojant šį analizės metodą gauti rezultatai leidžia ne tik daryti išvadas apie mokinių gebėjimus, mokymosi dėsningumus, mokymo metodus, žanrų požymius ir ribas ar vertimo problemas. Jie suteikia papildomą prieigą prie analizuojamų klausimų, atveria dar vieną jų analizės kryptį, leidžia konkrečia išraiška patvirtinti faktus, kurie iki tol galbūt buvo numanomi, tačiau sunkiau apčiuopiami. Pvz., čia minėtas Ai ir Lu (2013) tyrimas parodė labai aiškią skirtį tarp gimtakalbių ir negimtakalbių studentų rašinių, juos vertinant keturiais sintaksinio sudėtingumo aspektais. Autorių atliktas tyrimas parodė, kad negimtakalbių studentų rašiniai yra trumpesni, turi menkesnį kiekį subordinacijos ryšių, jų vartojamos struktūros yra paprastesnės. Tokio pobūdžio tyrimas labai aiškiai parodo konkrečias problemas, kurias mokymo metodologijų kūrėjai gali imtis spręsti. Lygiai taip pat galima paminėti ir Casal bei Lee (2019) tyrimą, kurio rezultatai atkreipia dėmesį į konkrečias sakinio struktūras, būdingas sėkmingiems ne gimtakalbių rašytiems akademiniams tekstams. Analizuodami vardažodinių frazių sudėtingumą, autoriai nustatė, kad didesnis tokių frazių sudėtingumas (ypač tais atvejais, kai modifkatoriais ėjo būdvardžiai, dalyviai ir adpozicijos) turi

reikšmingą ryšį su kokybiniu teksto vertinimu. Remiantis tokiomis išvadomis galima pateikti aiškias metodines gaires. Įdomiu ir tai, kaip sintaksinio sudėtingumo metodai taikomi vertimo studijose. Pvz., Liu ir Afzaal (2021) palygino originalius ir išverstus tekstus bei patvirtino, kad vertimai yra mažiau sintaksiškai sudėtingi už originalus, tačiau kartu išryškėjo ir gana pastebima žanro svarba. Ypač įdomi autorių įžvalgos apie originalo kalbos normų įtaką vertimų kalbai. Kaip ir kitais atvejais, sudėtingumo tyrimas padeda papildyti pasirinktas teorines prieigas, atrasti naujų jų aspektų.

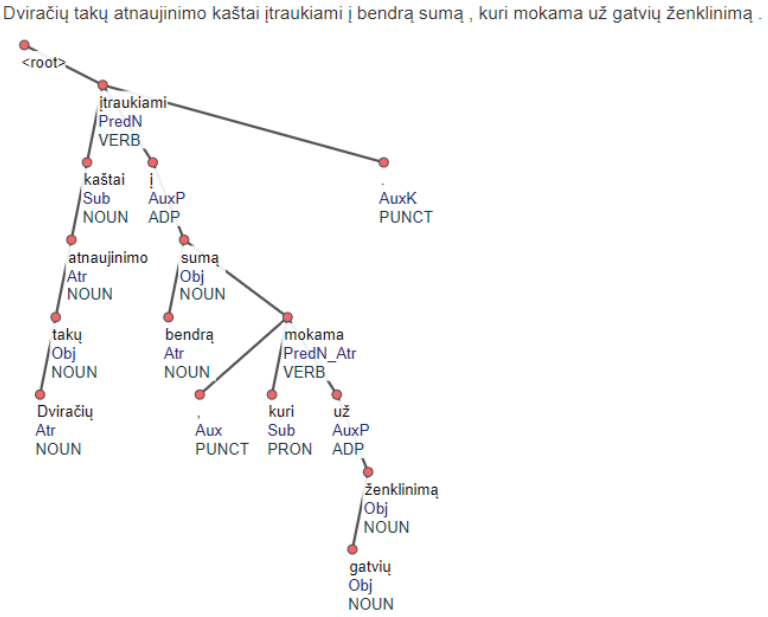
Čia paminėta tik dalis apie sintaksinį sudėtingumą rašiusių autorių, stengiantis įtraukti rastus naujausius tyrimus. Norint išsamiai apžvelgti tiek sudėtingumo sampratos paradigmą, tiek skirtingų sintaksinio sudėtingumo matavimų veiksmingumą ir panaudojimą reikėtų turbūt ne vieno, o kelių atskirų tyrimų ir straipsnių. Taip pat sudėtinga pateikti užtikrintą išvadą, kurios iš sudėtingumo matavimo metodikos krypties reikėtų laikytis ir kuris apibrėžimas ar samprata būtų tiksliausia ir veiksmingiausia. Todėl šiame darbe sintaksinis sudėtingumas yra traktuojamas kaip sukurto teksto ypatybė ir bendro teksto sudėtingumo aspektas, parodantis ne tik jo išmanumą, bet ir sunkumą. Sudėtingumo ribos nėra iš anksto apibrėžiamos, sudėtingumui arba paprastumui nesuteikiamos papildomos reikšmės, nė vienas nelaikomas geresnio, blogesnio, vertingesnio teksto požymiu, kadangi neįtraukiant gausios įvairių matavimo būdų ir jų lyginimo vertinamuoju aspektu visumos, imtis tokio vertinimo būtų netikslinga. Kaip jau minėta, šiame darbe pristatomas ir išbandomas savo procesu nuo anksčiau aprašytų sudėtingumo matavimo būdų besiskiriantis sintaksinio sudėtingumo analizės metodas – priklausomybės nuotolio nustatymas. Toliau kiek plačiau aptariamas šis matavimo būdas ir pristatoma tyrimo metodologija.

Priklausomybės nuotolis: tyrimo metodologija ir medžiaga

Tyrimo metodologija. Oya (2010) pasiūlė lingvistinį sudėtingumą matuoti remiantis grafų teorijos aspektais. Tokia prieiga grįsta jau minėta *dependency locality* teorija, kur sintaksinis sudėtingumas siejamas su atminties patiriama apkrova. Idėja atrodo labai paprasta – ilgesnė priklausomybės distancija išaugina bendrą sintaksinio sudėtingumo rodiklį, kadangi reikalaujama daugiau atminties resursų susiejant priklausomą žodį su tuo, nuo kurio priklauso. Tas pats Oya (2011) pristatė tyrimą, kuriame matavo studentų, besimokančiųjų anglų kalbos, sukurtų tekstų vidutinį priklausomybės nuotolį (toliau – VPN, ang. ADD – average dependency distance) ir pateikė preliminarias išvadas, kuriose teigiama, kad didesnė teksto apimtis žodžiais nebūtinai koreliuoja su didesniu VPN. Priklausomybės nuotolis kalbos sudėtingumui ar sunkumui pamatuoti buvo naudojamas dar iki Oya. Liu (2008) priklausomybės nuotolį naudojo lyginant kalbų sudėtingumą, o 2017 m. publikacijoje Liu et al. aprašė pastebėtą universalų polinkį mažinti priklausomybės nuotolį ir išryškėjusius tam tikslui pasitelkiamus sintaksinius modelius. 2018 m. publikuotame Lei ir Jockers straipsnyje dar kartą apžvelgiama priklausomybių nuotolio matavimo metodika. Pastebima, kad abejonių kelia priklausomybės nuotolio ir sakinio apimties santykis. Pagrįstai teigiama (tą pastebi ir Oya), kad sakinio ilgis gali turėti įtakos (tam tikrais išskirtiniais atvejais) priklausomybės nuotolio rodikliui. Dėmesys atkreipiamas į tuos atvejus, kai tiriami ne vienodo ilgio sakiniai. Anot autorių, tokiu būdu rezultatai gali būti netikslūs. Netikslumų gali kilti ir dėl sintaksinio medžio viršūnės (ang. *root*), kurios pozicijos svarba, anot autorių, nebuvo tinkamai įvertinta iki tol atliktuose VPN matavimuose. Jos svarbą galima pagrįsti argumentu, jog sintaksinio priklausomybių medžio viršūnę galima laikyti ir savotišku sakinio atskaitos tašku, nuo kurio priklauso tiek jo supratimas, tiek kūrimas. Minimo straipsnio autorių manymu, skaitymo ir teksto supratimo procese papildomi atminties resursai išnaudojami dar iki nustatant sakinio viršūnę, į kurią toliau atsiremia viso sakinio supratimas. Todėl sudarant sintaksinio sudėtingumo apskaičiavimo formulę, svarbu įtraukti ir šią dedamąją. Šiam tikslui autoriai pasiūlė formulę, kuria apskaičiuojamas modifikuotas vidutinis priklausomybės nuotolis (toliau – MVPN). Atliktas tyrimas ir skirtingų kalbų palyginimas parodė, kad MVPN nuo sakinio ilgio priklauso mažiau už VPN. Šiame straipsnyje pristatomi tiek MVPN, tiek VPN analizės duomenys.

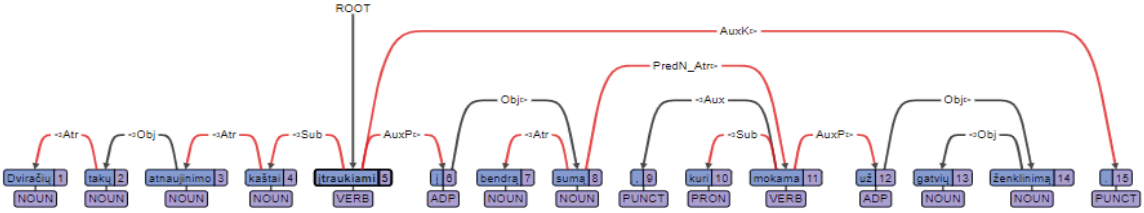
Vidutinis priklausomybės nuotolis sintaksiškai anotuotame tekstyne apskaičiuojamas gana paprastai. Tai leidžia ne tik greitai atlikti reikalingą duomenų analizę, bet ir patogiai ją pritaikyti tolimesniems tyrimams. Priklausomybės nuotolis apskaičiuojamas iš sintaksiškai anotuoto sakinio, įvertinant linijinę distanciją tarp dviejų žodžių, susietų priklausomybės ryšiu. Pirmame paveiksle pateiktas pavyzdys iš sintaksiškai anotuoto lietuvių kalbos tekstyno ALKSNIS. Kaip matyti pateiktame pavyzdyje, sintaksiškai anotuoti sakiniai pateikiami grafiškai medžio principu, taip vizualiai perteikiant priklausomybės ryšius. Šis atvaizdas sugeneruotas naudojant internetinį įrankį

CoNLL-U Viewer, nemokamai prieinamą internete. Tokios formos medyje viršūne laikomas hierarchiškai aukščiausias sakinio dėmuo (*įtraukiami*), prie kurio briaunomis prijungti kiti taškai (žodžiai ar skyrybos ženklai), taip nurodant jų sintaksinę priklausomybę.



1 pav. Sintaksiškai anotuoto lietuvių kalbos tekstyno ALKSNIS priklausomybių medis

Skaičiuojant VPN dėmesys pirmiausia sutelkiamas į linijiškumą, todėl čia labiau tinka kitoks grafo atvaizdavimo būdas, pavaizduotas antrame paveiksle[2].



2 pav. Sintaksiškai anotuoto lietuvių kalbos tekstyno ALKSNIS priklausomybės ryšiai

Pateiktame pavyzdyje (**2 pav.**) yra atvaizduotas tas pats sintaksiškai anotuotas sakinys su pažymėtais priklausomybės ryšiais. Čia taip pat matyti, kad priklausomybės distancija siejasi su tiek su hierarchine sakinio priklausomybių struktūra, tiek su jo linijiškumu. Kaip jau minėta, priklausomybės nuotolis parodo, kiek sakinyje esantis žodis nutolęs nuo žodžio, su kuriuo yra susietas priklausomybės ryšiu. Svarbu atkreipti dėmesį ir į tai, kad gali skirtis anotavimo formatai. ALKSNIS yra parengtas naudojant PML (*Prague Markup Language*)[3] formatą. Tas pats sakinys anotuotas dabar jau paplitusiu UD (*Universal Dependencies*)[4] formatu atvaizduojamas jau kiek kitaip.

[2] Šiam atvaizdui naudota nemokamai prieinama programa „UD Annotatrix". Nuorodas į šį ir kitus tokio tipo failaims apdoroti skirtus įrankius galima rasti adresu https://universaldependencies.org/.

[3] PML – sintaksiniam anotavimui ir priklausomybių medžių vizualizavimui bei redagavimui naudojamas lingvistinių duomenų formatas.

[4] UD – Skirtingoms kalboms skirta nuoseklaus gramatikos anotavimo sistema.

Skirtumai atsiranda dėl skirtingų anotavimo taisyklių ir principų. Abu formatai tokio pobūdžio tyrimui suteikia savų privalumų ir trūkumų. Pvz., UD iš esmės skirtas kiek galima labiau standartizuoti skirtingų kalbų anotavimą, žodžius priklausomybės ryšiais jungia tik su kitais reikšminiais žodžiais, viršūne niekada, skirtingai nuo PLM, negali būti skyrybos ženklas (norint juos įtraukti į VPN apskaičiavimo formulę reikalinga atlikti išsamesnę analizę, kuri peržengia šio darbo ribas). Iš vienos pusės tai labiau atitinka VPN skaičiavimo metodikos esmę, iš kitos, kaip teigia Lei ir Jockers (2018), UD formato reikalavimai gali lemti savotišką ryšių atvaizdavimo dirbtinumą. Funkcinių elementų neįtraukimas į grafą, nors ir pateisinamas siekiant kaip galima didesnio universalumo, gali lemti labiau semantinę, o ne sintaksinę viso medžio struktūrą. Šiame tekste toliau aprašoma tik PML formatu anotuotų ALKSNIO failų analizė, turint omenyje visus galimus šio formato trūkumus ir siekiant nustatyti galimybę ateityje juos taisyti ir atitinkamai koreguoti gautus rezultatus.

VPN formulė nėra sudėtinga. Skaičiavimai atliekami iš anotavimo metu parengtų CoNLL-U[5] formato failų, kuriuose atskirai pažymėtas kiekvieno atskiro vieneto priklausomybės nuotolis. Pirmoje lentelėje pateikiama išsami vieno apdoroto sakinio informacija.

1 lentelė Sintaksiškai anotuoto sakinio infromacija CoNLL-U formatu

# text = Dviračių takų atnaujinimo kaštai įtraukiami į bendrą sumą, kuri mokama už gatvių ženklinimą.					
1	Dviračių	NOUN	dkt.vyr.dgs.K.	2	Atr
2	takų	NOUN	dkt.vyr.dgs.K.	3	Obj
3	atnaujinimo	NOUN	dkt.vyr.vns.K.	4	Atr
4	kaštai	NOUN	dkt.vyr.dgs.V.	5	Sub
5	įtraukiami	VERB	vksm.dlv.neveik.es.vyr.dgs.V.	0	PredN
6	į	ADP	prl.G.	5	AuxP
7	bendrą	NOUN	dkt.mot.vns.G.	8	Atr
8	sumą	NOUN	dkt.mot.vns.G.	6	Obj
9	,	PUNCT	skyr.	11	Aux
10	kuri	PRON	įv.mot.vns.V.	11	Sub
11	mokama	VERB	vksm.dlv.neveik.es.mot.vns.V.	8	PredN_Atr
12	už	ADP	prl.G.	11	AuxP
13	gatvių	NOUN	dkt.mot.dgs.K.	14	Obj
14	ženklinimą	NOUN	dkt.vyr.vns.G.	12	Obj
15	.	PUNCT	skyr.	5	AuxK

Dėl patogumo pateikiamas supaprastintas variantas, be sakinio identifikavimo ir papildomų žymų informacijos, kuri tokiai analizei, bent jau šiame žingsnyje, neturi esminės reikšmės.

Skaičiuojant VPN svarbiausia informacija yra pirmame ir penktame stulpeliuose. Pirmame kiekvienas sakinio vienetas (žodžiai ir skyrybos ženklai) iš eilės sunumeruotas didėjančia tvarka. 5 stulpelyje pažymėta, prie kurio sakinio vieneto prijungtas konkretus žodis ar skyrybos ženklas. Nuliu pažymėta sakinio viršūnė arba ne viršūne esantis skyrybos ženklas. Kaip jau minėta, priklausomybės nuotoliu laikoma distancija tarp dviejų priklausomybės ryšiu sujungtų sakinio vienetų. 1 lentelėje pateiktame pavyzdyje, sakinio viršūne laikomas dalyvis *įtraukiami* yra 5 sakinio pozicijoje, subjektas *kaštai* 4. Jų priklausomybės nuotolis – 1. Tokiu atveju viso sakinio sudėtingumas apskaičiuojamas sudėjus visą priklausomybės nuotolį ir padalinus iš sakinio priklausomybės ryšių kiekio. Šiame darbe į formulę neįtraukiami ne viršūne esantys skyrybos ženklai. Konkrečiu 1 lentelėje esančiu atveju kablelis būtų apskaičiuojamas kaip turintis priklausomybės nuotolį lygų 0 ir neįtraukiamas į bendrą priklausomybės ryšių kiekį:

[5] CoNLL-U – gramatinių pažymų pateikimo forma.

$$VPN = \frac{1+1+1+1+1+1+2+1+3+1+1+2}{12} = 1{,}33$$

Atitinkamai viso teksto VPN apskaičiuojamas randant visų sakinių priklausomybės nuotolio sumą ir padalinant ją iš sakinių skaičiaus.

Kaip jau minėta, šitoje formulėje neatsižvelgiama į sakinio viršūnę ir galimą sakinio ilgio poveikį. Šios problemos sprendimui Lei ir Jockers (2020) pasiūlo sudėtingumą matuoti naudojant tokią formulę:

$$MVPN = abs\ (ln\ (\frac{VPN}{\sqrt{Viršūnės\ nuotolis * sakinio\ ilgis}}))$$

Čia MVPN yra modifikuotas sakinio priklausomybės nuotolis, *viršūnės nuotolis* – viršūnės pozicija sakinyje, *sakinio ilgis* – priklausomybės ryšių (grafo briaunų) kiekis (kaip ir anksčiau, neskaičiuojant skyrybos ženklų). Pagal šią formulę pateikto sakini MVPN būtų apskaičiuojamas taip:

$$MVPN = abs\ (ln\ (\frac{1{,}33}{\sqrt{5 * 12}})) = 1{,}76$$

Šiame darbe, naudojant abi pristatytas formules, siekiama nustatyti sintaksiškai anotuoto tekstyno sintaksinio sudėtingumo rodiklius. Nustatomas kiekvieno sakinio ir kiekvieno teksto sintaksinio sudėtingumo rodiklis, tarpusavyje lyginamos tekstyno dalys, pristatomos išryškėjusios pirminės sudėtingumo lygmenų nustatymo galimybės, aptariamas galimas palyginimas su kitais sintaksinio sudėtingumo nustatymo būdais, aptariami išryškėję didžiausi nuokrypiai, pateikiamos tolimesnių tyrimų gairės ir rekomendacijos.

Tyrimo medžiaga. Tyrimui naudojama medžiaga iš sintaksiškai anotuoto lietuvių kalbos tekstyno ALKSNIS. Jame pateikti 3643 sintaksiškai anotuoti sakiniai, peržiūrėti ir sutvarkyti rankiniu būdu, todėl duomenys naudojami jų niekaip nemodifikuojant. Tiesa, yra kelios išimtys, kurias būtina paminėti.

ALKSNIS sudarytas iš penkių dalių (bendroji periodika, grožinė literatūra, mokslinė literatūra, specialioji periodika ir administraciniai tekstai). Dėl administracinių tekstų specifikos ir dokumentų rengimo reikalavimų, tekstų sakiniai tekstyne skaidomi papunkčiui arba skirties tašku pasirenkant kabliataškį. Todėl ALKSNIO administraciniuose tekstuose vienas sakinys suskaidytas ir pateikiamas kaip keli atskiri sakiniai, pvz.

> *Šiam tikslui įgyvendinti numatomos šios Lietuvos kultūros politikos kaitos gairės: // 1) įtvirtinti kultūrą kaip strateginę valstybės raidos kryptį, teikiant prioritetą kultūros politikai; // 2) reformuoti ir demokratizuoti kultūros valdymą, plėtojant kultūros savireguliaciją;* <...>

Taip pat iš analizuojamų sakinių pašalinti dokumentų skyrių pavadinimai, datos, pasirašymo vietos, autorių nuorodos. Atlikus šiuos pakeitimus analizuotos medžiagos apimtis atrodo taip:

2 lentelė Tyrimo medžiagos apimtis

	Grožinė literatūra	Bendroji periodika	Specialioji periodika	Administraciniai tekstai	Mokslinė literatūra
Tekstai	18	34	7	8	9
Sakiniai (analizuoti)	639	697	671	716	643

Iš viso analizuoti 3366 sakiniai. Verta pastebėti, kad medžiagos apimtis nėra didelė, tačiau, atsižvelgiant į darbo pobūdį ir tikslus, laikoma pakankama.

Galiausiai reikia paminėti apie išskirtinius atvejus, kuriems kol kas nepavyko rasti tinkamo sprendimo, tačiau jų kiekis sąlyginai mažas, kad darytų lemiamą įtaką tyrimo pasirinkimui ir metodų taikymui. Analizuojant sakinius pasirinktu metodu, neapskaičiuojamas vienanarių ar benarių sakinių sudėtingumas, kadangi trūkstant priklausomybės ryšių, formulės *sakinio ilgis* vietoje atsiranda 0. Tokių sakinių sudėtingumas šiame darbe vertinamas nuliu. Nors tekstyne jų kiekis mažas (1,5 proc.), ilgainiui bus reikalinga rasti tikslesnį įvertį

Rezultatai

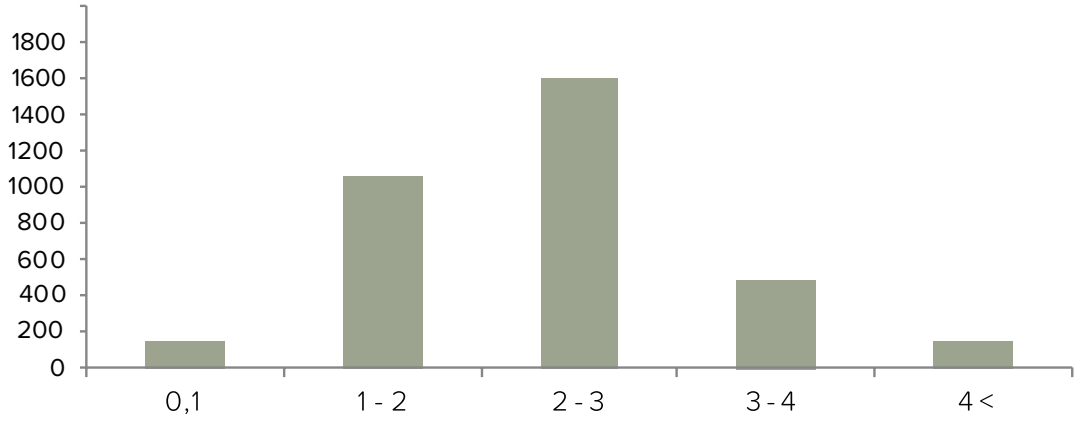
Vidutinis priklausomybės nuotolis Apskaičiavus viso tekstyno sakinių sintaksinį sudėtingumą, pirmiausia buvo apžvelgti bendri visų sakinių duomenys. Nustatytas vidutinis sakinio sintaksinio sudėtingumo rodiklis – 2.332. Mažiausias rodiklis – 0. Tai jau minėti vienanariai ar benariai sakiniai. Atmetus juos, sintaksinio sudėtingumo rodiklis varijuoja tarp 1.0000 ir 8.5109 (aptiktas vienas išskirtinis atvejis, kai VPN buvo virš 31, tačiau toks rezultatas atsiranda dėl iš eilės einančių vienarūšių sakinio elementų, skirtingų produktų pavadinimų ir skyrybos ženklų, kurie neįskaičiuojami į formulę, tačiau padaugina bendrų sakinio pozicijų kiekį. Laikoma, kad VPN čia išpūstas dirbtinai. Didžiausią VPN turintis sakinys yra iš administracinių tekstų, t. y. *Smulkiojo ir vidutinio verslo plėtros įstatymo*[6] (šiame ir kituose tekstyno pavyzdžiuose skyrybos ženklai yra atitraukti ir atvaizduoti taip, kaip būtų užrašyti sintaksinio medžio analizės programoje, kadangi kiekvienas iš jų medyje būtų pažymėti atskiru tašku su savo briauna). Vis dėlto laikoma, kad šis atvejis kyla dėl iškreiptų ir nenuoseklių duomenų. Įdomu ir tai, kad šio sakinio MPN nėra aukštas – 1.2145. Akivaizdu, kad konkretus pavyzdys iškrenta iš bendro konteksto dėl gana paprastos priežasties. Kiti panašūs sakiniai tekstyne yra suskaidyti, kiekvieną atskirą punktą nagrinėjant kaip atskirą sakinį. Panašiai išsiskiria ir kiti sakiniai, kurių VPN yra neįprastai aukštas. Pvz., administracinės dalies teksto *Veiklos_ataskaita* 116 sakinys:

> *Tarptautiniuose renginiuose ir darbo grupėse skaityti šie pranešimai : „Duomenų apsauga Lietuvoje“, „E-valdžios plėtra Lietuvoje : asmens tapatybės nustatymas ir duomenų apsauga“, „Tiesioginė rinkodara ir privatumo problemos“, „Vaizdo stebėjimas ir duomenų apsauga“, „Duomenų apsaugos naujovės Lietuvoje “ , „Lietuvos duomenų apsaugos pažanga telekomunikacijose“, „Valstybinės duomenų apsaugos inspekcijos patirtis taikant Direktyvos 2002 58 EB 5 straipsnio 2 dalyje numatytą išimtį“, „Duomenų apsaugos problemos atliekant vaizdo stebėjimą“, „Įgyvendinimo grupės tikrinimas privačiame sveikatos draudimo sektoriuje 2006“ .*

Šio sakinio sintaksinio sudėtingumo rodiklis 7.4857, o MVPN tik 1.0071. To priežastis taip pat tampa akivaizdi žvelgiant į priklausomybių grafą. ALKSNYJE vienarūšiai dėmenys jungiami sujungiamuoju ryšiu. Kadangi vienarūšių dėmenų daug, sujungiamojo ryšio rodiklis (kuriuo šiuo atveju laikomas paskutinis sakinio kablelis) yra nutolęs nuo apibendrinamojo dėmens, prie jo jungiamų dėmenų nuotolis taip pat atitinkamai didelis, taip išpučiant bendrą sakinio VPN. Tai puikus pavyzdys, kaip sakinio ilgis ir jo dėmenų kiekis gali išauginti sintaksinio sudėtingumo rodiklį ir kodėl yra reikalinga modifikuoti matavimo formulę. Privalu laikytis kaip galima didesnio anotavimo nuoseklumo ir atsižvelgti į galimus duomenų netikslumus, ypač dirbant su administraciniais tekstais.

Siekiant įvertinti sudėtingumo pasiskirstymą ir vėliau analizuoti sakinius pagal sudėtingumo lygmenis, nuspręsta apžvelgti sakinių pasiskirstymą pagal nustatytus rodiklius. Preliminariai išskirtos tokios ribos: 0 ir 1, 1–2, 2–3, 3–4, >4. Pirmoje diagramoje pavaizduota, kaip sakiniai pasiskirstė pagal šiuos lygmenis.

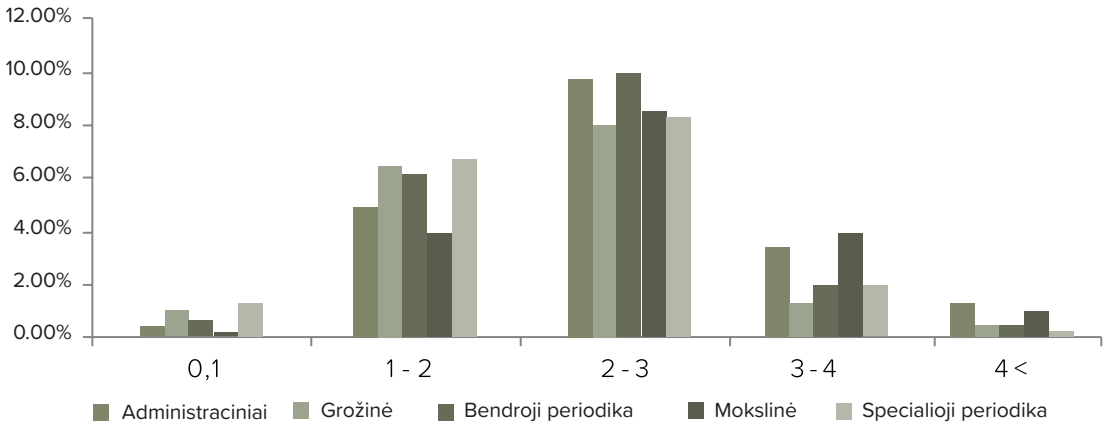
[6] Smulkiojo ir vidutinio verslo subjektams gali būti taikomos šios valstybės paramos formos : 1) mokesčių lengvatos (jei jos nustatytos įstatymų), rinkliavų lengvatos; 2) teisės aktų nustatyta tvarka finansinė parama: lengvatinių paskolų teikimas, labai mažų paskolų teikimas, dalinis ar visiškas palūkanų dengimas, garantijų teikimas, kreditų draudimas, rizikos kapitalo investavimas, tam tikrų išlaidų (steigimo, tyrimų, garantijų mokesčių, kreditų draudimo įmokų, sertifikavimo (registravimo), atitikties įvertinimo ir kitų) kompensavimas, subsidijos darbo vietoms kurti; 3) viešųjų paslaugų verslui teikimas verslo inkubatoriuose, verslo informacijos centruose, mokslo ir technologijų parkuose ir kituose juridiniuose asmenyse, kurių steigimo dokumentuose nustatytas šių paslaugų teikimas; 3 punkto redakcija nuo 2011 m. sausio 1 d. : 3) viešųjų paslaugų verslui teikimas viešojoje įstaigoje „Eksportuojančioji Lietuva“, verslo inkubatoriuose, verslo informacijos centruose, mokslo ir technologijų parkuose ir kituose juridiniuose asmenyse, kurių steigimo dokumentuose nustatytas šių paslaugų teikimas; 4) Vyriausybės ar savivaldybių nustatytos kitos paramos formos.



1 diagrama Sakinių pasiskirstymas tekstyne pagal sudėtingumą

Kaip galima matyti lentelėje, beveik pusė iš visų tekstyno sakinių (47,9 proc.) yra įvertinti sintaksinio sudėtingumo rodikliu tarp 2 ir 3, trečdalis (30,5 proc.) – tarp 1 ir 2, o 14 procentų – tarp 3 ir 4. Kaip jau minėta, kraštinius atvejus šiuo metu įvertinti yra sunku. Nors iš pirmo žvilgsnio sunku kalbėti apie išryškėjančius sudėtingumo laiptus, remiantis tokiu pasiskirstymu galima parengti eksperimentinius tyrimus, kurių metu būtų galima įvertinti sukurtų ir specialiai tokiam tyrimui sumodeliuotų tekstų poveikį skaitymo laikui, suvokimui, tirti akių judesius ir pan. Turint tokių papildomų tyrimų duomenis, būtų galima tiksliau atsakyti, kiek reikšmingas yra skirtumas tarp sudėtingumo rodiklių.

2 diagramoje pateiktas skirtingų tekstyno dalių sakinių pasiskirstymas pagal kategorijas.



2 diagrama Skirtingų tekstyno dalių sakinių pasiskirstymas pagal sudėtingumą

Iš šios diagramos matyti, kad didžiausiais sudėtingumo rodikliais pasižymi administraciniai ir moksliniai tekstai. Čia didžiausia dalis nuo visų tekstyno sakinių buvo įvertinti 4 arba daugiau. To priežastys iš dalies jau aptartos. Mažiausius sudėtingumo rodiklius daugiausia turi grožinė literatūra ir specialioji periodika. To taip pat galima tikėtis, kadangi grožinėje literatūroje gausiau trumpų, vos kelių žodžių sakinių.

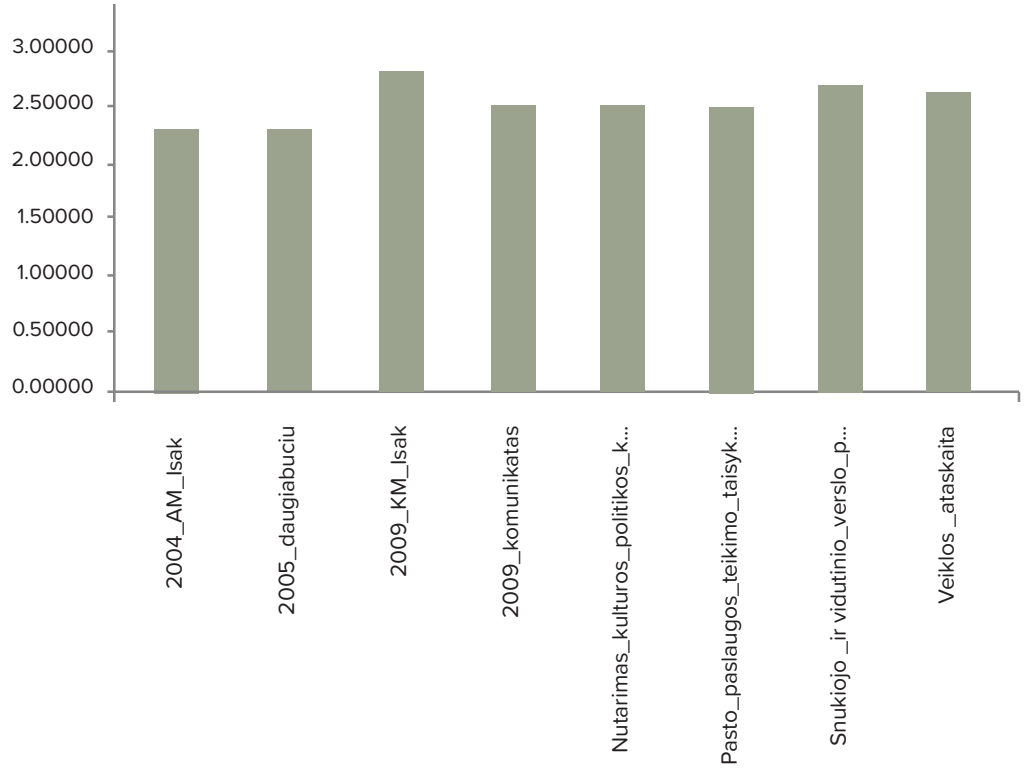
Taip pat buvo lyginti viso teksto sudėtingumo rodikliai, gauti bendrą sakinių VPN padalinus iš sakinių skaičiaus. 76 tekstyno tekstų VPN rodikliai pasiskirstė tarp 1.7376 ir 3.8022. Vidutinis tekstyne analizuotų tekstų rodiklis – 2.2954. Trečioje lentelėje pavaizduota, kaip sudėtingumo rodikliai pasiskirstė tarp skirtingų tekstyno dalių:

3 lentelė Vidutinis sakinio ir teksto sudėtingumas skirtingose tekstyno dalyse yrimo medžiagos apimtis

	vieno sakinio VPN	vieno teksto VPN
Grožinė literatūra	2.1198	2.1704
Specialioji periodika	2.1295	2.1701
Bendroji periodika	2.2568	2.2332
Mokslinė literatūra	2.6369	2.6467
Administraciniai tekstai	2.5500	2.5556

Įdomu tai, kad žvelgiant į grožinės ir specialiosios periodikos dalį, skirtumų beveik nematyti. Beveik vienodi čia ir vieno sakinio, ir vieno teksto vidutiniai rodikliai. Kaip ir buvo galima tikėtis, aukščiausius įverčius turi mokslinės literatūros ir administraciniai tekstai. Nors reikalinga išsamesnė statistinė analizė, pirminiai duomenys rodo, jog skirtumas tarp rezultatų (lyginant grožinės ir mokslinės literatūros visų sakinių ir tekstų VPN) yra statistiškai reikšmingas ($p<0,01$).

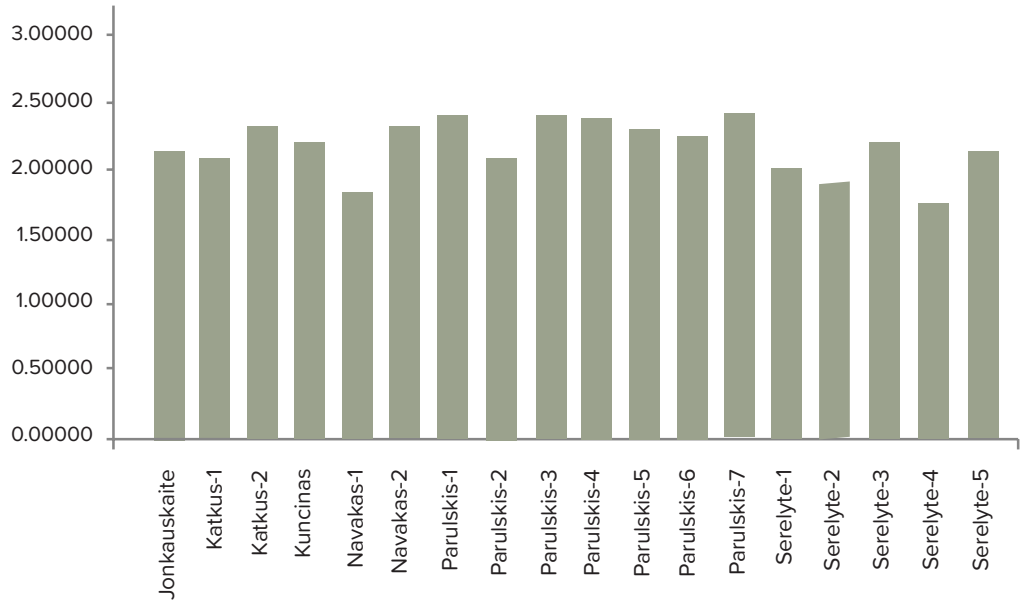
Administracinių tekstų sudėtingumas pasiskirstė taip:



3 diagrama Administracinių tekstų sintaksinio sudėtingumo pasiskirstymas, pagal VPN

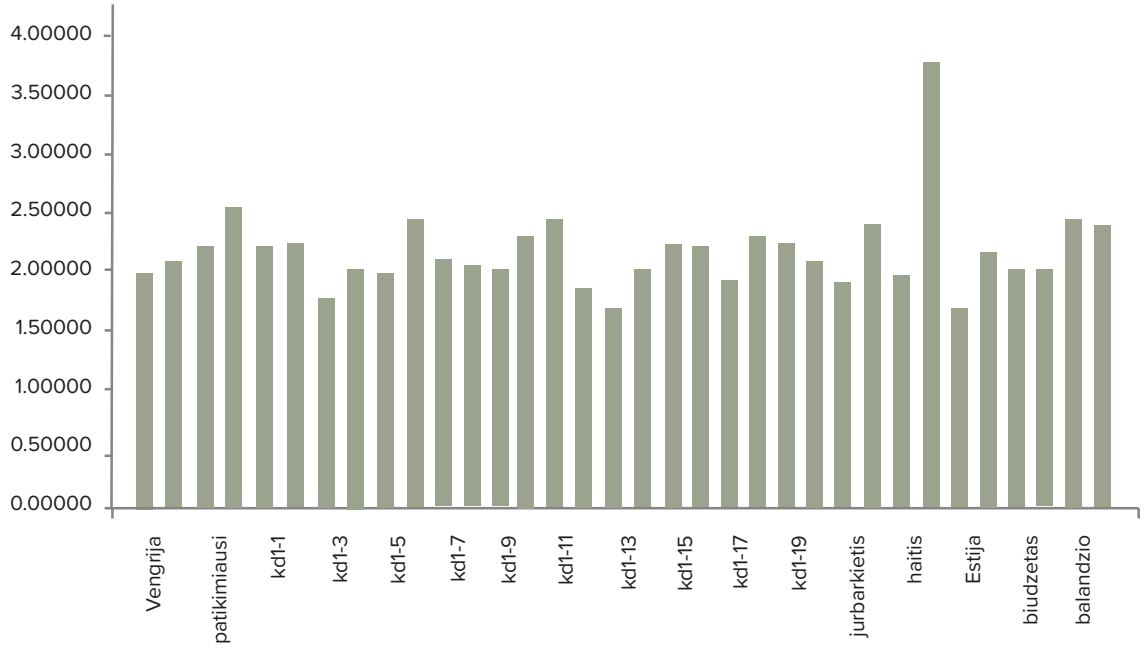
Įdomu tai, jog šiuo metodu sudėtingiausiu įvertintas tekstas aukščiausią įvertį taip pat turėjo ir sintaksinio sudėtingumo tyrime, kurį atliko Kalinauskaitė (2019). Tuo metu su ankstesne tekstyno versija atliktame tyrime buvo analizuota mažiau tekstų, tačiau atitinkamų tekstų pasiskirstymas pagal sudėtingumo lygmenis panašus.

Panašiai kaip Kalinauskaitės darbe, sintaksiniam sudėtingumui apskaičiuoti naudojant VPN, grožinės literatūros tekstuose taip pat matyti gana ženklus verčių svyravimas. Reikalinga išsamesnė analizė to priežastims nustatyti, tačiau viena galimų hipotezių – skirtingas autorių stilius, sintaksiniam sudėtingumui turintis daugiau įtakos, negu nusistovėjusios žanro normos. Net ir paviršutiniškai žvelgiant į tekstyne atrinktų autorių tekstų duomenis, galima pastebėti skirtumų (pvz., vidutinis pateiktų Šerelytės ir Parulskio tekstų VPN (2.0098 prieš 2.3258)). Užtikrintų teiginių ir išvadų daryti negalima, tačiau tai galėtų būti indikacija tęsti pasirinktos metodologijos analizę ir apsvarstyti skirtingas metodo panaudos galimybes.



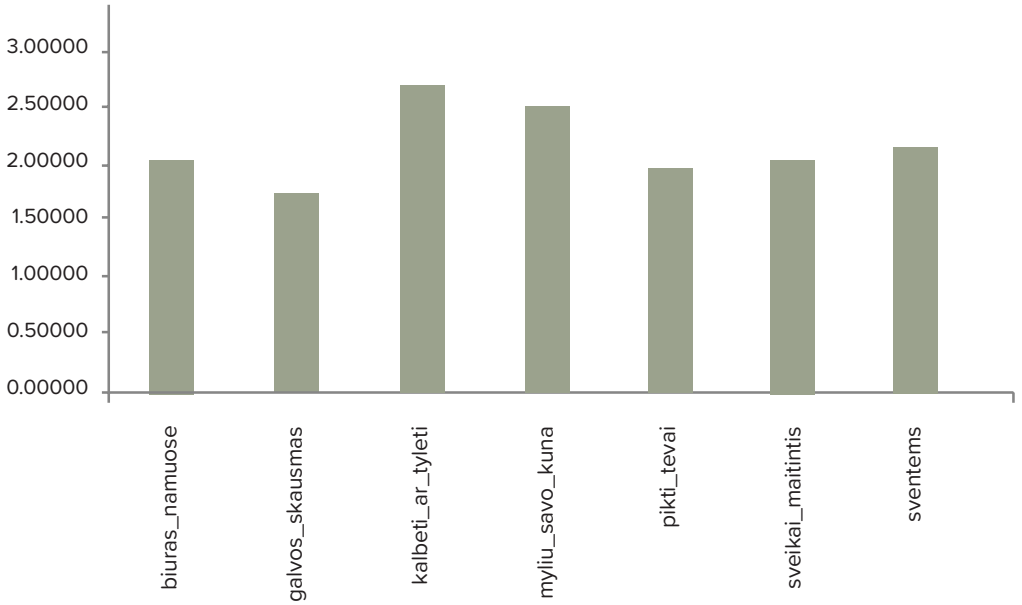
4 diagrama Grožinių tekstų sintaksinio sudėtingumo pasiskirstymas, pagal VPN

Bendrosios periodikos VPN taip pat išsiskiria gana ženkliu pasiskirstymu. Rezultatai svyruoja nuo 1.7376 iki 3.8022. Tai taip pat galima pateisinti rašymu apie skirtingas temas ir tuo, jog tekstai skirti skirtingoms auditorijoms. Ryškiai didžiausią įvertį turinčio teksto duomenis iškreipia du dirbtinais laikytini rodikliai (virš 7 ir virš 31), kurie gauti iš sakinių, su dideliu kiekiu vardijamų vienarūšių pavadinimų, panašiai, kaip anksčiau minėtu atveju.



5 diagrama Bendrosios periodikos tekstų sintaksinio sudėtingumo pasiskirstymas, pagal VPN

Specialiosios periodikos tekstų VPN pasiskirstymas pavaizduotas 6 diagramoje:



6 diagrama Bendrosios periodikos tekstų sintaksinio sudėtingumo pasiskirstymas, pagal VPN

Šios tekstyno dalies rezultatai svyruoja tarp 1.7430 ir 2.4709. Kaip jau minėta, dėmesį patraukia specialiosios periodikos ir grožinės literatūros VPN artumas. Ypač turint mintyje, kad ir pagal priskirtą žanrą, ir publikavimo vietą šie tekstai turėtų būti artimesni bendrajai publicistikai. Tačiau į tokius panašumus reikia žvelgti atsargiai, dėl šiame darbe jau ne kartą paminėto ALKSNIO minuso – mažos apimties. Vis dėlto tai atveria galimybes kalbėti apie sintaksinio sudėtingumo ir VPN matavimo galimybes žanro ribų ir žanro ypatybių tyrimuose.

Kalbant apie mokslinės literatūros tekstų VPN rodiklį taip pat galima pastebėti santykinį nuoseklumą, buvusį būdingą ir administraciniams tekstams. Tekstų VPN svyruoja tarp 2.4983 ir 2.9740. Taip pat matyti, kad visos vertės yra artimos 2.5000. Svarstant apie to priežastis pirmiausia reikėtų įvertinti tiek mokslo, tiek administraciniams tekstams būdingą šabloniškumą, informacijos pateikimo pobūdį. Gana ilguose sakiniuose atsiranda didžiuliai priklausomybės distancijos nuotoliai, kuriuos lemia tokio stiliaus tekstams būdingos konstrukcijos. Pvz., sakinio *Vertinant smurto aktus taip pat buvo pasitelkti du nepriklausomi žurnalistų etikos inspektoriaus tarnybos ekspertų grupės nariai* . VPN yra 2.9333 (jei pagal susitarimą 3 laikytume sintaksiškai aukštą sudėtingumą turinčio sakinio riba, tai čia ji būtų beveik pasiekta, nors sakinio sandara, žvelgiant iš jau minėtos Kalinauskaitės darbo apibrėžčių, nelaikytina sudėtinga). Tai daugiausia lemia subjekto pozicija sakinyje ir jį nuo subjekto (ir kitų atributų) skiriantys pažyminiai. Tai taip pat galima traktuoti kaip anksčiau iškeltos minties pagrindą – sakinį sudėtingu galima laikyti dėl to, jog jame reikalaujama nemažai atminties resursų, kol, skaitant iš eilės, yra užbaigiamas priklausomybės ryšys.

Modifikuotas vidutinis priklausomybės nuotolis. Iki šiol aptariant rezultatus nebuvo atsižvelgta į sakinių ilgį ar sakinio viršūnės poziciją. Kaip jau minėta, šiuos elementus į sudėtingumo skaičiavimo formulę įtraukti galima, taip minimizuojant galimą rezultatų išsikreipimą ar priklausomybę nuo sakinio ilgio. Kaip to pavyzdį galima dar kartą priminti jau aptartą administracinio teksto sakinį, kurio modifikuotas priklausomybės nuotolis nebuvo aukštas. Dar vienas akivaizdus to pavyzdys jau minėta anomalija iš bendrosios publicistikos teksto *GMO*, kurios VPN įvertis virš 31[7]. Jos sudėtingumą vertinant MVPN gaunamas rezultatas – 0.8143, kuris yra žemiau šios tekstyno dalies MVPN vidurkio.

[7] Taip pat 22 pavadinimų augaliniai aliejai: „Brolio“, „Lankų“, „Sodžiaus“, „Kolumbo“, „Tėviškės“, „Augalinis aliejus“, „Dolores“, „Maxima“, „Optima linija“, „Perla“, „Karolina“, „Žemaičio“, „Aukselis“, „Saulutė“, „Omili“, „Huilor“, „Oilio“, „Vitela“, „Luccija“, „Jasmine“, „Caroli“, „Zitos sojų aliejus“.

Nustatytas vidutinis visų tekstyno sakinių MVPN – 1.2580. Vertės svyruoja nuo 0 iki 3.3175. Kaip ir VPN, 0 rezultatas atsiranda dėl vienanarių ar benarių sakinių, kai formulėje dėl priklausomybės ryšių nebuvimo įrašomas 0. Tolimesniuose darbuose būtų reikalinga tokiems atvejams apsvarstyti atskirą įvertį. Atmetus 0 mažiausi įverčiai prasideda nuo 0.0119. Taip įvertintas specialiosios periodikos teksto *biuras_namuose* sakinys:

Vis gi, nepaisant ekonominių ir ekologinių privalumų, darbas namuose yra nemažas iššūkis.

Pažvelgus į šį sakinį atidžiau, kyla du su MVPN susiję klausimai: tikslumo ir patikimumo. Ar jis gali tiksliai atspindėti lietuvių kalbos teksto sudėtingumą, ir ar jo atskirų dedamųjų poveikis nėra neproporcingai per didelis. To paties sakinio VPN – 3.2000, o tai preliminariu vertinimu galima traktuoti kaip gana aukštą sintaksinio sudėtingumo rodiklį, nors dėl toliau minimų priežasčių ir šio įverčio tikslingumas ir tikslumas taip pat gali kelti abejonių. Situacija čia, regis, kebli. Bendru požiūriu sutikti, jog toks sakinys gali turėti tokį mažą sintaksinio sudėtingumo įvertį – sunku. Tai nėra itin trumpas sakinys, priklausomybės ryšiais susieti elementai nutolę vienas nuo kito, predikatas sakinio pabaigoje. Dar daugiau, čia matomas sujungimas teksto lygmeniu – sakinys pradedamas sujungiamuoju jungtuku, kuris rodo sąsają su ankstesniu sakiniu. Kalbant apie sudėtingumą šio darbo kontekste, reikėtų labai rimtai apsvarstyti, ar tokio tipo sakiniams nevertėtų įvesti papildomo koeficiento, kadangi galima daryti prielaidas, jog tokiais atvejais turėtų būti reikalaujama dar daugiau atminties išteklių, o tekstas, kuriame tokių sakinių gausiau, galėtų būti sudėtingesnis. Tačiau kol kas visa tai prielaidos, kurioms patikrinti reikalingi tolesni tyrimai. Grįžtant prie konkretaus aptariamo pavyzdžio galima matyti, kad sakinio viršūnės pozicijos reikšmė yra ženkli. Jei pirmas sakinys būtų šiek tiek pakeistas, pašalintas *visgi*, o visa kita palikta nepakeista, sakinio MVPN ženkliai šoktelėtų – iki 1.2747, kas yra daugiau už bendrą visų sakinių vidurkį. Tokių atvejų tekstyne yra ne vienas. Turbūt lemiamas tokių pokyčių ir rezultatų veiksnys – anotavimas ir jo susitarimai. Panašu, kad norint išspręsti čia paminėtas problemas reiktų persvarstyti anotavimo principus, atsižvelgiant į atvejus, galinčius turėti įtakos galutiniams rezultatams. Nors šis metodas turėtų minimalizuoti sakinio ilgio poveikį, panašu, kad tuo galima suabejoti. Atidžiau pažvelgus į aukštą sudėtingumo rodiklį gavusius sakinius matyti, kad jie pasižymi didesne priklausomybės ryšių gausa (yra ilgesni), o sakinio viršūnė yra gerokai pasislinkusi į antrąją sakinio pusę. Nors tokia ir buvo deklaruota formulės kūrėjų intencija, panašu, kad sakinio ilgio faktorius čia išlieka. Juo labiau, kad palyginus priklausomybės ryšių kiekį ir MVNP vienoje tekstyno dalių (grožinės literatūros dalyje), nustatyta statistiškai reikšminga ($p<0,001$) koreliacija ($r = 0,648$). Vis dėlto remiantis tuo, kas buvo išdėstyta anksčiau, nereikėtų skubėti atmesti šios formulės kaip nepatikimos.

Aptikus aptartus netikslumus, nuspręsta kol kas giliau duomenų neanalizuoti, pateikiant tik apibendrintą informaciją apie skirtingas tekstyno dalis.

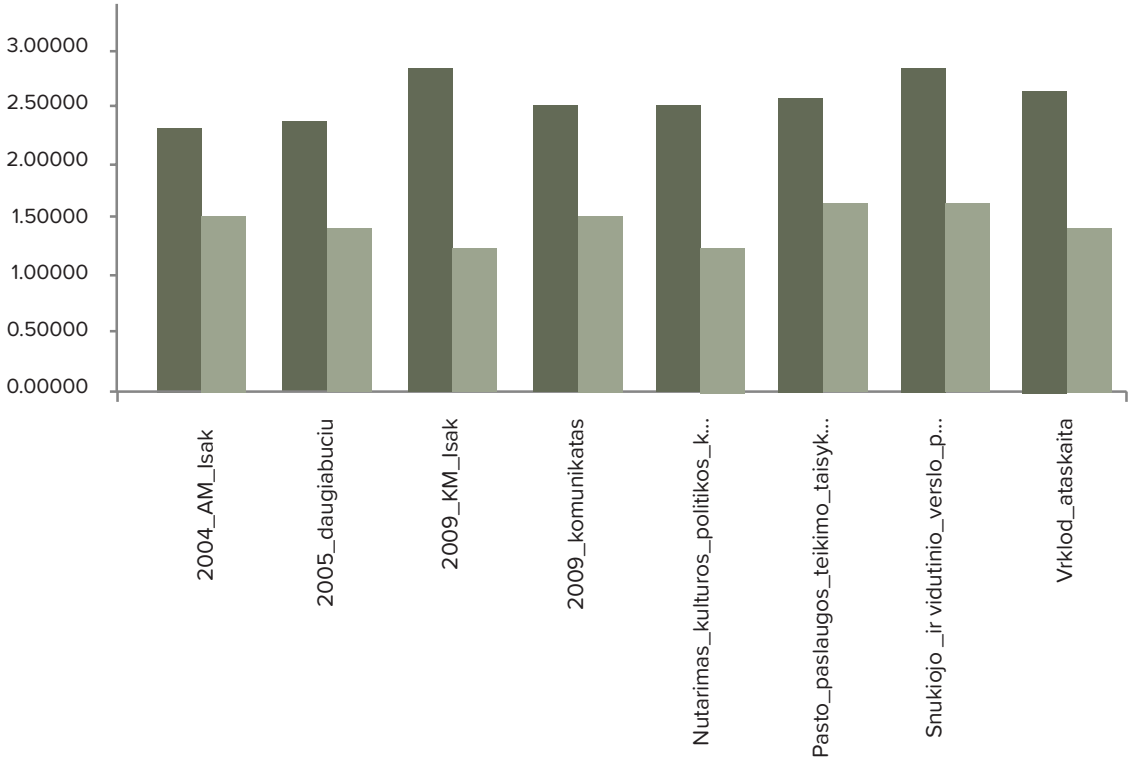
4 lentelė Modifikuotas vidutinis sakinio ir teksto sudėtingumas skirtingose tekstyno dalyse

	vieno sakinio VPN	vieno teksto VPN
Grožinė literatūra	1.0616	1.1396
Specialioji periodika	1.1111	1.0848
Bendroji periodika	1.2963	1.3484
Mokslinė literatūra	1.359	1.3424
Administraciniai tekstai	1.44	1.4323

Iš ketvirtos lentelės matyti, kad sudėtingiausi sakiniai, pagal dabartinės formos MVPN, administraciniuose ir moksliniuose tekstuose. Sudėtingiausiais laikytini administraciniai tekstai, o bendroji periodika ir mokslinė literatūra taip pat viršija bendrą visų tekstų vidurkį (1.2830). Tai iš dalies atitinka VPN rezultatus, nors gana ženkliai skiriasi mokslinės literatūros įvertis. Įdomu tai, kad lyginant su Kalinauskaitės atliktu tyrimu (kadangi sudėtingumas ten analizuotas visai kitaip, palyginimas tik paviršinis), administraciniai tekstai pasižymi didžiausiu sudėtingumu, o jos darbe – mažiausiu. Grožinė literatūra atvirkščiai – tiek MVPN, tiek VPN rodo žemesnį sudėtingumą, o ankstesniame tyrime grožinės literatūros tekstai pasirodė turintys didžiausią sudėtingumo rodiklį.

Palyginus atskirų tekstyno dalių tekstų sintaksinio sudėtingumo rodiklį, gautą apskaičiavus MVPN, nustatyti nuo VPN besiskiriantys rezultatai. Kaip pavyzdys čia pateikiama tik vienos tekstyno dalies – administracinių tekstų sudėtingumo rodiklio palyginimas (**7 diagrama**). Čia taip pat matyti gana tolygus pasiskirstymas, tačiau reikėtų atkreipti dėmesį į tekstą *2009_KM_Isak,* kurio VPN yra aukščiausias, o MVPN – žemiausias. Panašu, kad tokį skirtumą lemia jau

aptartos priežastys. Į minimo teksto rodiklius pažvelgus atidžiau, matyti pasikartojantys atvejai, kai ilgų, gana aukštą VPN sakinių viršūnė yra pačioje sakinio pradžioje, kas lemia žemą MVPN. Tokių sakinių pradžiose predikatai *pakeičiu, išdėstau, tvirtinu*. Konkretus pavyzdys – sakinys[8], kurio VPN virš 4, o MPN tik 0.7040 – žemiau įprasto vidurkio. Panašius skirtumus galima stebėti ir žvelgiant į kitų tekstyno dalių tekstų sudėtingumo rodiklius.



7 diagrama Vidutinio priklausomybės nuotolio ir modifikuoto priklausomybės nuotolio palyginimas administraciniuose tekstuose

Apibendrinimai ir išvados

Šiame darbe pristatyti du sintaksinio sudėtingumo matavimo būdai. Siekta bendrai aptarti esminius jų principus ir išbandyti naudojant lietuvių kalbos sintaksiškai anotuoto tekstyno ALKSNIS duomenis. Bene pagrindinė šio bandymo išvada labai paprasta – laukia dar daug darbo, kad šiais sintaksinio sudėtingumo matavimo būdais būtų galima pasitikėti ir juos naudoti automatinei ar pusiau automatinei analizei atlikti. Atlikta analizė išryškino ne tik pagrindinius trūkumus, kuriuos reikia apsvarstyti ir pašalinti prieš remiantis šiais matavimais, bet ir tam tikrus turimų duomenų apdorojimo niuansus, kuriems taikoma metodologija pasirodė esanti jautri. Taip pat iškilo ir papildomų klausimų apie pačią sakinio ir teksto sintaksinio sudėtingumo sąvoką.

Pirma problema – duomenų pateikimo nenuoseklumas. Norint, kad sintaksinio sudėtingumo matavimai naudojant VPN ir, ypač MVPN, būtų tikslūs, būtina turėti vieningą ir nuoseklią, konkrečiai tam pritaikytą anotavimo schemą. Aptarti ALKSNIO atvejai parodė, kad kai kurie esminiais laikytini anotavimo principai gali paveikti rezultatus, ypač tais atvejais, kai šių principų nėra laikomasi.

Taip pat verta apsvarstyti kitą anotavimo sistemą. Jau minėta, kad priklausomybės nuotolį galima apskaičiuoti

[8] Pakeičiu Valstybės biudžeto lėšų skyrimo programos „ Meno kūrybos plėtra ir sklaida Lietuvoje ir užsienyje “ priemonės „ Vykdyti projekto „ Vilnius – Europos kultūros sostinė 2009 “ parengiamuosius darbus “ projektų daliniam finansavimui tvarkos aprašą , patvirtintą Lietuvos Respublikos kultūros ministro 2008 m . rugsėjo 22 d . įsakymu Nr . ĮV-459 „ Dėl Valstybės biudžeto lėšų skyrimo programos „ Meno kūrybos plėtra ir sklaida Lietuvoje ir užsienyje “ priemonės „ Vykdyti projekto „ Vilnius – Europos kultūros sostinė 2009 “ parengiamuosius darbus “ projektų daliniam finansavimui tvarkos aprašo , paraiškos , sutarties ir ataskaitų formų patvirtinimo “ (Žin . , 2008 , Nr . 109-4179) :

pagal UD. Nors šia sistema anotuotų lietuvių kalbos išteklių yra (juos galima rasti puslapyje universaldependencies.com), kol kas jie nėra patikimi, kadangi galima pastebėti nemažai anotavimo netikslumų ir nenuoseklumų. Todėl tolimesni darbai apimtų tokias užduotis – anotavimo sistemos parinkimą ir/ar pritaikymą, išsamią duomenų analizę ir palyginimą su rezultatais, išryškėjusiais ALKSNIO analizėje. Taip pat svarbu surasti būdą, kaip vertinant sintaksinį sudėtingumą apimti ne tik ryšius sakinio viduje, bet ir sąsajas tarp sakinių, kurių svarba taip pat išryškėjo žvelgiant į kai kuriuos iš aptartų rezultatų. Tai galėtų būti papildomas koeficientas ar formulės modifikavimas, atliktas po išsamesnės analizės.

Kita užduotis, kuri laukia norint toliau plėtoti priklausomybės nuotolio ir sintaksinio sudėtingumo tyrimą – nustatyti sintaksinio sudėtingumo lygmenis. Kadangi pati metodologija yra susijusi su teksto supratimo procesais, manoma esant tikslinga sudėtingumo lygmenų nustatymui įtraukti eksperimentinius tyrimus, kurie padėtų išryškinti aiškiau apibrėžtas ribas tarp sudėtingais ir mažiau sudėtingais laikytinų sakinių.

Galiausiai būtina reikšmingai plėsti tyrimo apimtis. Analizuojant iki 30 vienam stiliui ar vienai tekstyno daliai priskirtų tekstų, neįmanoma daryti patikimų išvadų nė apie sudėtingumo požymius skirtingo stiliaus tekstuose, nė, tuo labiau, apie lietuvių kalbos tekstų sudėtingumą apskritai. Juolab nebūtų galima lietuvių kalbos lyginti su kitomis kalbomis, net jeigu galima numanyti, jog tokie lyginimai būtų prasmingi ateityje.

Literatūros sąrašas

1 Ai, H., & Lu, X. (2013). A corpus-based comparison of syntactic complexity in NNS and NS university students' writing. Studies in Corpus linguistics, 59, 249-264. John Benjamins Publishing Company. https://doi.org/10.1075/scl.59.15ai

2 Ansarifar, A., Shahriari, H., & Pishghadam, R. (2018). Phrasal complexity in academic writing: A comparison of abstracts written by graduate students and expert writers in applied linguistics. Journal of English for Academic Purposes, 31, 58-71. https://doi.org/10.1016/j.jeap.2017.12.008

3 Barrot, J. S. (2015). Comparing the linguistic complexity in receptive and productive modes. GEMA Online Journal of Language Studies, 15(2), 65-81. https://doi.org/10.17576/gema-2015-1502-05 https://ejournals.ukm.my/gema/article/view/7038/3337

4 Beers, S. F., & Nagy, W. E. (2011). Writing development in four genres from grades three to seven: Syntactic complexity and genre differentiation. Reading and Writing: An Interdisciplinary Journal, 24(2), 183-202. https://doi.org/10.1007/s11145-010-9264-9

5 Biber, D., Gray, B., Staples, S., & Egbert, J. (2020). Investigating grammatical complexity in L2 English writing research: Linguistic description versus predictive measurement. Journal of English for Academic Purposes, 46. https://doi.org/10.1016/j.jeap.2020.100869

6 Borillo, J. M. (2000). The degree of grammatical complexity in literary texts as a translation problem. Benjamins Translation Library, 32, 65-74. https://doi.org/10.1075/btl.32.09gar

7 Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. Dimensions of L2 Performance and Proficiency, 21-46. DOI: 10.1075/lllt.32.02bu https://doi.org/10.1075/lllt.32.02bul

8 Casal, J. E., & Lee, J. J. (2019). Syntactic complexity and writing quality in assessed first-year L2 writing. Journal of Second Language Writing, 44, 51-62. https://doi.org/10.1016/j.jslw.2019.03.005

9 Chen, M., & Zechner, K. (2011). Computing and evaluating syntactic complexity features for automated scoring of spontaneous non-native speech. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, 722-731. Available at: https://www.aclweb.org/anthology/P11-1073

10 Díez-Bedmar, M. B., & Pérez-Paredes, P. (2020). Noun phrase complexity in young Spanish EFL learners' writing: Complementing syntactic complexity indices with corpus-driven analyses. International Journal of Corpus Linguistics, 25(1), 4-35. https://doi.org/10.1075/ijcl.17058.die

11 Evans, R., & Orăsan, C. (2019). Identifying signs of syntactic complexity for rule-based sentence simplification. Natural Language Engineering, 25(1), 69-119. https://doi.org/10.1017/S1351324918000384

12 Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. Cognition, 68(1), 1-76. https://doi.org/10.1016/S0010-0277(98)00034-1

13 Jagaiah, T., Olinghouse, N. G., & Kearns, D. M. (2020). Syntactic complexity measures: Variation by genre, grade-level, students' writing abilities, and writing quality. Reading and Writing, 33(10), 2577-2638. https://doi.org/10.1007/s11145-020-10057-x

14 Jiang, K., Biber, D., & Gray, B. (2016) Grammatical complexity in Academic English: Linguistic change in writing. Applied Linguistics, 37. https://doi.org/10.1093/applin/amw035

15 Kalinauskaitė, D. (2019). Lietuvių kalbos tekstų informatyvumo nustatymas. [Doctoral dissertation]. Vytauto Didžiojo universitetas.

16 Karlsson, F. (2014). Complexity in linguistic theorizing. The Mental Lexicon, 9. https://doi.org/10.1075/ml.9.2.01kar

17 Lahuerta Martinez, A. C. (2017). Syntactic complexity in secondary level English writing: Differences among writers enrolled on bilingual and non-bilingual programmes. Porta Linguarum, 28, 67-80. https://doi.org/10.30827/digibug.54003

18 Larsson, T., & Kaatari, H. (2020). Syntactic complexity across registers: Investigating (in) formality in second-language writing. Journal of English for Academic Purposes, 45. https://doi.org/10.1016/j.jeap.2020.100850

19 Lei, L., & Jockers, M. L. (2020). Normalized dependency distance: Proposing a new measure. Journal of Quantitative Linguistics, 27(1), 62-79. https://doi.org/10.1080/09296174.2018.1504615

20 Kortmann, B., Szmrecsanyi, B. (2012). Linguistic Complexity: Second Language Acquisition, Indigenization, Contact. de Gruyter, Walter GmbH & Co.https://www.degruyter.com/view/title/37350 https://doi.org/10.1515/9783110229226

21 Liu, H. (2008). Dependency Distance as a Metric of Language Comprehension Difficulty. https://doi.org/10.17791/JCS.2008.9.2.159

22 Liu, H., Xu, C., & Liang, J. (2017). Dependency distance: A new perspective on syntactic patterns in natural languages. Physics of Life Reviews, 21. https://doi.org/10.1016/j.plrev.2017.03.002

23 Liu, K., & Afzaal, M. (2021). Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. PloS one, 16(6). https://doi.org/10.1371/journal.pone.0253454

24 Lowrey, T. M. (1998). The effects of syntactic complexity on advertising persuasiveness. Journal of Consumer Psychology, 7(2), 187-206. https://doi.org/10.1207/s15327663jcp0702_04

25 Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL Quarterly, 45(1), 36-62. https://doi.org/10.5054/tq.2011.240859

26 Ma, Q., & Wang, X. (2019). What is language complexity? Macrolinguistics, 7, 1-29. https://doi.org/10.26478/ja2019.7.11.1

27 Miestamo, M., Sinnemaki, K., & Karlsson F. (2008). Language complexity. Studies in Language Companion Series, 94. John Benjamins Publishing Company. https://benjamins.com/catalog/slcs.94 https://doi.org/10.1075/slcs.94

28 Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. Applied Linguistics, 24(4), 492-518. https://doi.org/10.1093/applin/24.4.492

29 Ortega, L. (2015). Syntactic complexity in L2 writing: Progress and expansion. Journal of Second Language Writing, 29, 82-94. https://doi.org/10.1016/j.jslw.2015.06.008

30 Oya, M. (2011). Syntactic dependency distance as sentence complexity measure. Proceedings of the 16th Conference of Pan-Pcific Association of Applied Linguistics, 313-316.

31 Pallotti, G. (2015). A simple view of linguistic complexity. Second Language Research, 31(1), 117-134. https://doi.org/10.1177/0267658314536435

32 Rescher, N. (1998). Complexity: A Philosophical Overview. Transaction Publishers.

33 Shadloo, F., Ahmadi, H. S., & Ghonsooly, B. (2019). Exploring syntactic complexity and its relationship with writing quality in EFL argumentative essays. Topics in Linguistics, 20(1), 68-81. https://doi.org/10.2478/topling-2019-0005

34 Stauder, A., & Ustaszewski, M. (2020). Syntactic complexity as a stylistic feature of subtitles. Studia Translatorica, 11, 177. https://doi.org/10.23817/strans.11-13

35 Steger, M., & Schneider, E. W. (2012). Complexity as a function of iconicity: The case of complement clause constructions in

New Englishes. In Linguistic Complexity (pp. 156-191). De Gruyter. https://www.degruyter.com/view/book/9783110229226/10.1515/9783110229226.156.xml https://doi.org/10.1515/9783110229226.156

36 Szmrecsanyi, B., Purnelle, G. (Ed.), Fairon, G. (Ed.), & Dister, A. (Ed.) (2004). On Operationalizing Syntactic Complexity. Le poids des mots. Proceedings of the 7th International Conference on Textual Data Statistical Analysis, 2, 1032-1039. http://www.benszm.net/omnibuslit/Szmrecsanyi2004.pdf

37 Kortmann, B., & Szmrecsanyi, B. (2012). Linguistic Complexity: Second Language Acquisition, Indigenization, Contact. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110229226

38 Thongyoi, K., & Poonpon, K. (2020). Phrasal complexity measures as predictors of EFL university students' English academic writing proficiency. REFLections, 27(1), 44-61.

39 Wu, X., Mauranen, A., & Lei, L. (2020). Syntactic complexity in English as a lingua franca academic writing. Journal of English for Academic Purposes, 43. https://doi.org/10.1016/j.jeap.2019.100798

40 Yin, S., Gao, Y., & Lu, X. (2021). Syntactic complexity of research article part-genres: Differences between emerging and expert international publication writers. System, 97. https://doi.org/10.1016/j.system.2020.102427

Abstract

Vytautas Ožeraitis. Analysis of Syntactic Complexity in the Annotated Lithuanian Language Corpus by the Method of Dependency Distance

Syntactic complexity is a feature common to all languages and is generally described as an assessment of the sophistication, elaborateness, length, and patterns of a sentence (or text) and its elements. In Lithuanian, syntactic complexity is not widely analyzed. Studies of syntactic complexity are problematic due to the unstable definition of the term and the abundance of different methods for calculating it. This article presents the study of syntactic complexity in the syntactically annotated Lithuanian corpus ALKSNIS, using the syntactic dependency distance method, which is based on the Dependency Locality theory. The article introduces the concept of syntactic complexity, presents the principles of its research, their relevance, and discusses the results of syntactic complexity in the corpus, advantages, and disadvantages of the chosen method. This study aims to supplement the field of the syntactic complexity analysis of the Lithuanian language.

For the analysis of syntactic complexity, two measures are used: the mean dependency distance and the modified mean dependency distance. The study analyzes corpus data, determines the syntactic complexity of individual sentences and texts. A detailed analysis reveals both the shortcomings of the methods used and their dependence on an accurate and consistent annotation scheme. Analyzing the data, the need to include linkages between sentences into syntactic complexity formulas becomes apparent. The position of the sentence vertex included in the modified mean dependence distance formula has been found to potentially distort the results, hence the study calls for further refinement of the formula. The boundaries of the complexity of sentences and texts identified in the present study are indicative, hence further qualitative analysis and experiments are needed to define them with greater precision.

Apie autorių

VYTAUTAS OŽERAITIS

Vytauto Didžiojo universitetas, Lietuva

Mokslinių interesų kryptys

Tekstynų lingvistika, skaitmeninė lingvistika, konceptualiųjų metaforų raiška, diskurso analizė

Adresas

J. Bilūno g. 18, 53527 Garliava, Kauno r., Lietuva

El. paštas vytautas.ozeraitis@vdu.lt