Lithuanian Journal of Statistics
Lietuvos statistikos darbai
Vol. 61, 2022, pages 32–45
https://doi.org/10.15388/LJS.2022.31988

Vilnius University Press

Skiemenų statistikos taikymas atskiriant poeziją nuo prozos

Gediminas Murauskas, Marijus Radavičius

Taikomosios matematikos institutas, Vilniaus universitetas
Naugarduko g. 24, LT-03225 Vilnius, Lietuva

El. paštas: gediminas.murauskas@mif.vu.lt; marijus.radavicius@mif.vu.lt

Įteiktas 2021 gruodžio 31; pataisytas 2022 lapkričio 11; publikuotas 2022 gruodžio 20

Santrauka. Straipsnio tikslas – sukonstruoti klasifikatorių, kuris pagal trumpas teksto ištraukas galėtų atskirti poeziją nuo prozos ir kurį kuo mažiau įtakotų atskirų autorių stilius ir kūrinių turinys. Todėl apmokant klasifikatorių naudojama tik informacija apie tekstų skiemenis, nes pastarieji atspindi kalbos fonetines savybes ir mažiau negu žodžiai yra susiję su tekstų turiniu. Tyrimas remiasi suskaitmenintų grožinės literatūros kūrinių bibliotekos http://ebiblioteka.mkp.emokykla.lt tekstais. Jų pagrindu sudarytas ir apmokytas klasifikatorius atskirdamas testinius 100 žodžių ilgio poezijos ir prozos tekstų fragmentus darė mažiau negu 5% klaidų.

Raktiniai žodžiai: logistinė regresija; automatinis skiemenavimas; kryžminė patikra; apmokymas; klasifikavimo klaida
AMS: 62H30, 91G20, 68T50

1 Įvadas

Spartus kompiuterių mokslo ir technikos vystymasis ir vis platesnis jų naudojimas sąlygojo kiekybinių metodų, tuo pačiu ir duomenų analizės bei statistikos, skverbimąsi į kalbotyrą. Statistinė duomenų analizė taikoma daugelyje kalbotyros krypčių: žodžių pasiskirstymų ir tekstų nehomogeniškumo tyrimuose, sprendžiant tekstų klasifikavimo (pagal žanrą, stilių ir pan.) ir autoriaus identifikavimo uždavinius ir kt. (žr., pvz., [22, 24, 5, 15] ir ten cituojamą literatūrą). Šiuose darbuose, kaip ir daugumoje kompiuterinės lingvistikos tyrimų (išskyrus tyrimus, skirtus šnekamosios kalbos generavimui ir atpažinimui), remiamasi *žodžių* statistika. Kuo įdomūs *skiemenys*?

- Dauguma bazinių žodžių, paprastai tai – tarnybiniai arba funkciniai žodžiai, yra vienskiemeniai, taigi, sutampa su skiemenimis.
- Skiemuo sudaro šnekamosios kalbos pagrindą, jos vientisą elementą (vienetą).
- Bet svarbiausias motyvas domėtis skiemenimis yra noras ištirti skiemenavimo panaudojimo galimybes mažiau įprastiems kalbotyros uždaviniams negu žodžių kėlimas, kirčiavimas ir šnekamosios kalbos sintezė ar atpažinimas [7, 8, 12].

Skiemens pagrindą, jo būtiną dedamąją sudaro balsis ar balsių junginys. Kebliau nustatyti skiemens ribas.

Yra įvairios skiemens ribų teorijos ir apibrėžimai, pavyzdžiui: Pakerys (2007) manė, kad skiemuo – šnekamosios kalbos atkarpa, kurios garsai sudaro minimalų artikuliacinį, akustinį ir funkcinį vienetą; skiemuo yra vienetas, padedantis atskleisti, kaip garsai yra derinami ir išdėstomi kalboje (čia remtasi [10, 11]).

Sprendžiant techninį automatinio lietuvių kalbos skiemenavimo uždavinį paprastai naudojamas fonologinio skiemens apibrėžimas [9, 2, 10, 11, 19, 12]. Fonologinio skiemens pradžios riba eina ten, kur prasideda didžiausia vidinio priebalsių junginio dalis, savo struktūra sutampanti su atitinkamu žodžio pradžios junginiu [3]. Abejonių šiuo racionaliu skaidymo į skiemenis principu kelia ženklūs pirmojo ir antrojo bei pirmojo ir vidinių fonologinių skiemenų sudėties skirtumai [6, 4]. Kadangi ilgesnes priebalsių samplaikas sunkiau ištarti, natūralu skaidyti priebalsių junginius ir skiemens ribas apibrėžti taip, kad skiemens pradinio ir galinio priebalsių junginio struktūra derėtų su atitinkamomis potencialiai galimomis žodžio pradžios ir pabaigos priebalsių struktūromis, bet būtų kuo trumpesnės, lengviau ištariamos ir identifikuojamos. Šią taisyklę ir naudosime skiemenuodami žodžius.

Raškinis ir Kazlauskienė savo darbe [19] rašo: „Tai, kad nėra patikimos metodikos, nustatančios objektyvias, fonetinėmis ypatybėmis paremtas skiemenų ribas, rodo, jog kalbai svarbesnis skiemenų skaičius, o ne tikslios ribos tarp jų“. Lietuviškų žodžių skiemenų skaičiaus pasiskirstymo dėsningumus tyrė ir matematikai. Merkytė ir Kalinka (1968) [18] skiemenų skaičiaus skirstinio aprašymui pritaikė *Fux* modelį ir pasiūlė metodą to modelio nežinomiems parametrams statistiškai įvertinti. Taip pat buvo pastebėta, kad skiemenų ir fonemų skaičiaus tarpusavio priklausomybę lietuvių kalbos žodžiuose gana gerai aprašo tiesinės regresijos modelis [17].

Taikomoji tekstų ir kalbos analizė paprastai remiasi duomenimis apie žodžių ir jų junginių pasitaikymo tekstyne dažnius. Lyginant žodžius su skiemenimis pastarieji daug mažiau yra susiję su tekstų turiniu, bet išsaugo tam tikrą informaciją apie tekstų skambesį (fonetinę informaciją). Be to, kaip jau buvo minėta, gana didelė dalis funkcinių ir dažnai vartojamų žodžių (prielinksniai, įvardžiai, prieveiksmiai, jungtukai, jaustukai ir pan.) yra vienskiemeniai. Vadinasi, skiemenų vartosena atspindi ir kai kuriuos teksto morfologinius bei sintaksinius aspektus.

Šiame darbe šias skiemenų ypatybes bandoma pritaikyti konkrečiam uždaviniui: atskirti poeziją nuo prozos nepaisant atskirų autorių ypatumų ir tekstų turinio.

Sprendžiant šį uždavinį susiduriama su originalių didelės apimties tekstų automatinio skiemenavimo problema, ypač jeigu tas skiemenavimas nėra standartinis. Deja, kalbininkų ir kompiuterinės lingvistikos specialistų sukurti ir palaikomi automatinio skiemenavimo įrankiai tam nėra tinkami, nes (a) yra susieti su konkrečiu tekstynu ir jo infrastruktūra, kuri suteikia gana ribotas, faktiškai tik standartines ir menkai automatizuotas priemones tyrimui reikalingo tekstyno sudarymui, (b) jų skiemenavimo metodika naudojasi žodžių morfologija ir kalbininkų ekspertiniu darbu.

Todėl buvo panaudotos universalios standartinės procedūros, skirtos lietuviškų ir nelietuviškų žodžių skiemenavimui ir kėlimui į kitą eilutę (*hyphenation*) [16, 23, 20, 13, 14]. Jos buvo adaptuotos skiemenavimo uždaviniui ir patobulintos.

Tyrime naudojamas suskaitmenintų tekstų rinkinys,[1] sudarytas iš 5–8 kl. mokiniams rekomenduojamų lietuvių autorių ir išverstų į lietuvių kalbą užsienio autorių grožinės literatūros kūrinių (romanai, apsakymai, poemos, eilėraščiai, pjesės ir kt.). Tekstų rinkinį sudarė 80 kūrinių, kurie parašyti 63-ų autorių; iš viso tekstuose buvo 2567290 žodžių, iš jų – 206453 skirtingi (žodžių formos). Toliau šį rinkinį trumpai vadinsime *tekstynu*.

Antrajame skyrelyje detalizuojama darbe taikyta skiemenavimo procedūra. Trečiajame skyrelyje aprašoma duomenų rinkinių, skirtų klasifikatoriaus, atskiriančio poeziją nuo prozos, apmokymui ir testavimui, sudarymo procedūra. Rinkinius sudaro nagrinėjamų tekstų 100 iš eilės einančių žodžių fragmentai. Ketvirtajame skyrelyje pateikiami skiemenų ir su skiemenavimu susijusias žodžių savybes reprezentuojantys kintamieji, naudojami klasifikatoriaus apmokyme. Dviejų etapų klasifikatoriaus apmokymo procedūra aprašyta penktajame skyrelyje, o jo patikros testiniais duomenimis rezultatai aptariami šeštajame skyrelyje. Pabaigoje pateikti rezultatų apibendrinimai ir išvados.

2 Skiemenavimas

Daugiau nei prieš 30 metų Frankas Liangas [16] pasiūlė šablonais pagrįstą skiemenavimo algoritmą. Šį algoritmą įgyvendino programoje *patgen*, kuri yra standartinio TeX paketo dalis. Lietuviškų žodžių skiemenavimo (tiksliau, žodžių kėlimo) šablonus sukūrė Vytautas V. Statulevičius ir Yannis Haralambous [23]. Šie skiemenavimo šablonai yra naudojami OpenOffice programose su MYSpell. Sigitas Tolušis iš VTeX šiuos šablonus adaptavo skirtingoms koduotėms ir palaikymui TeXLive paketuose.[2]

Tekstų skiemenavimui vienas iš šio straipsnio autorių parašė *Python* programą, kuri naudoja *pyphen* funkcijų biblioteką [13, 14] ir lietuviškų žodžių skiemenavimo šablonų rinkinį [23, 20].

Žodžių kėlimo taisyklės skiriasi nuo skiemenavimo remiantis fonologiniu skiemens apibrėžimu. Jos daug mažiau ribojančios, bet draudžia, pavyzdžiui, eilutėje palikti vieną raidę. Dėl to (ir ne vien dėl to) minėtą lietuviškų žodžių skiemenavimo šablonų rinkinį teko atnaujinti. Skiemenavimo klaidos buvo taisomos rankiniu būdu siekiant, kad po skiemenavimo procedūros gauti skiemenys tenkintų žemiau išvardintas sąlygas 1–2:

1. Taisyklingų lietuviškų žodžių skiemenyse nėra balsių, atskirtų viena nuo kitos priebalsiais.
2. Priebalsių grupės skiemenyse turi lietuvių kalbai būdingą struktūrą.

   Priebalsiai skirstomi į pučiamuosius (s, z, š, ž, f, h, jiems priskiriamas ir dviraidis ch; ši klasė žymima simboliu S), sprogstamuosius (p, b, t, d, k, g ir afrikatos c, č, dz, dž; ši klasė žymima simboliu T) ir sklandžiuosius (l, m, n, r, v, j; žymima

[1] Tekstai paimti iš laisvai prieinamos skaitmeninės bibliotekos http://ebiblioteka.mkp.emokykla.lt/. Autoriai dėkoja šio rinkinio sudarytojams už galimybę susipažinti su skirtais jaunimui lietuvių ir užsienio autorių grožinės literatūros kūriniais ir panaudoti juos lietuvių kalbos tyrimams.

[2] Remiamasi svetainėje https://ltex.lt/apie-mus/ pateiktu aprašymu.

simboliu R). Pagal fonologijos teoriją, lietuviškų žodžių skiemenys gali prasidėti tik STR, ST, SR ar TR tipo daugianariais priebalsių junginiais: STR trinarę grupę galima skaidyti į dvinares, bet nekeičiant priebalsių klasių tvarkos. Analogiškai ir simetriškai lietuviškų žodžių skiemenys gali baigtis ne ilgesniais kaip trinariai priebalsių junginiais, kurie yra grupės RTS(K) skaidiniai, išlaikant priebalsių klasių tvarką [9]. Pastarojoje grupėje klasei K priklauso priebalsiai „k“ ir „t“. Klasės K atvejis „k“ pasitaiko tik liepiamosios nuosakos veiksmažodžių paskutiniame skiemenyje, o atvejis „t“ – bendraties veiksmažodžių sutrumpintose galūnėse. Todėl skiemenuojant žodžius nuo jų pradžios galima apsiriboti RTS tipo skiemens pabaigos priebalsių grupe ir jos skaidiniais ([19, 73 psl.], cf. [9, 3 psl.]).

3 Tyrimo imties sudarymas

Šiame skyrelyje aprašoma, kaip iš tekstyno kūrinių yra sudaromi duomenys klasifikatoriaus apmokymui ir testavimui.

Visus 80 tekstyno kūrinių suskaidome į iš eilės einančius ir nesikertančius fragmentus, sudarytus iš 100 iš eilės einančių žodžių. Nustatant fragmentų dydį remiamasi žodžių, o ne skiemenų skaičiumi, nes skiemenų vartosena priklauso nuo skiemens vietos žodyje, be to, žodžių skaičius fragmente nesikeičia keičiant skiemenavimo taisykles.

Uždavinys yra sudaryti ir apmokyti klasifikatorių, kuris, naudodamas tik informaciją apie nagrinėjamame fragmente sutinkamus skiemenis bei to fragmento žodžių bendras su skiemenavimu susijusias charakteristikas, pakankamai patikimai atskirtų poezijos fragmentus nuo fragmentų, paimtų iš prozos.

Poezijos kūriniams yra priskirti eilėraščiai ir poezija.

Prozos kūriniams yra priskirti romanai, apysakos, apsakymai, novelės, esė.

Neklasifikuoti kūriniai: nei poezijai, nei prozai nepriskirtos dramos, tragedijos, taip pat poemos (pagal žanrą jos yra artimos poezijai), bei pasakos, sakmės, legendos, padavimai ir kiti tautosakos kūriniai (jie pagal žanrą yra artimesni prozai, bet gali turėti ir poezijos elementų).

1. Patogumo dėlei yra sukuriamas klasės požymis, pavadintas *poezija*, kuris aprašo tris klases: *poezija* $= 1$ tada ir tik tada, kai tekstas, jo fragmentas ar elementas yra paimtas iš poezijos kūrinio, *poezija* $= -1$, kai tekstas, jo fragmentas ar elementas yra paimtas iš prozos kūrinio ir *poezija* $= 0$ likusiais atvejais, t.y., kai kūrinys, jo fragmentas ar elementas nėra nei poezija, nei proza.

   Pastarosios klasės fragmentai yra *pagalbiniai*. *Pagrindiniai* tyrimo duomenys yra sudaryti iš pirmųjų dviejų klasių fragmentų.

2. Testavimo aibę sudaro trijų tipų fragmentai. Tuos tipus nusako žymė *use*.

   Atsitiktinai paimta apie 25% *pagrindinių* kūrinių ir jų visi fragmentai yra priskirti testavimo aibei. Tokiu būdu parinkti fragmentai sudaro *originaliąją testavimo aibę*. Ją nusako požymis *use* $= -2$.

   Likusiųjų *pagrindinių* kūrinių fragmentai yra paskirstomi tarp apmokymo ir testavimo tokiu būdu. Imant iš eilės kūrinio fragmentus, kas trečias fragmentas yra priskiriamas *bendrajai testavimo aibei* ir jam suteikiama žymė *use* $= -1$, o abu fragmentai iš kiekvienos tarpinių fragmentų poros yra priskiriami *apmokymo aibei* su atitinkama žyme *use* $= 1$.

Visi *pagalbinių* duomenų fragmentai (fragmentai su klasės požymiu *poezija* = 0) sudaro *pagalbinę testavimo aibę*. Juos nusako žymė *use* = 0.

3. Duomenų, skirtų apmokymui ir testavimui, apimtys 100 žodžių dydžio fragmentais ir jų procentine išraiška pateiktos 1 lentelėje.

   (a) Poezijos kūriniai yra gana skirtingo dydžio ir sudaro gana mažą *pagrindinių* duomenų dalį (apie 4.82%). Matyt, todėl atsitiktinai *originaliajai testavimo aibei* priskirtų poezijos fragmentų dalis yra 19.78% (124 iš 627), pastebimai mažesnė už 25%. *Apmokymo* ir *bendroji testavimo aibė* sudarė atitinkamai 54.23% ir 26%.
   (b) Prozos kūrinių fragmentams atitinkamos proporcijos yra 31.33%, 45.84%, 22.83%.

   Skaičiuojant tekstų, skirtų apmokymui ir testavimui, apimtis skiemenimis gaunamos labai panašios proporcijos kaip ir skaičiuojant fragmentais (žr. 1 lentelę).

1 lentelė. Apmokymo ir testavimo imčių dydis, išreikštas fragmentais ir skiemenimis.

Imties tipas	Žanras	Fragmentų skaičius (%)	Skiemenys (%)
Mokymo imtis ($use = 1$)	Poezija	340 (54.23%)	54.32%
	Proza	5689 (45.84%)	45.78%
Bendroji testavimo ($use = -1$)	Poezija	163 (26.00%)	26.06%
	Proza	2828 (22.83%)	22.84%
Originalioji testavimo ($use = -2$)	Poezija	124 (19.78%)	19.62%
	Proza	3880 (31.33%)	31.38%

Pastaba. Požymis *use* nusako tris nesikertančias fragmentų, skirtų testavimui, aibes: *bendrąją*, *originaliąją* ir *pagalbinę*. Naudojant tokiu būdu sudarytas tris testavimo aibes ir apmokius klasifikatorių galima patikrinti, kokiu mastu klasifikatorius yra nepriklausomas nuo kūrinio turinio ir autoriaus, bei įvertinti poetiškumo laipsnį *pagalbinio testavimo* rinkinio kūriniams, t.y., kūriniams, kurie nebuvo priskirti nei poezijai, nei prozai.

4 Sukurti aiškinantieji kintamieji ir požymiai

Skiemuo yra šnekamosios kalbos bazinis vienetas. Tačiau skiemenų vartoseną įtakoja ne vien tik žodžių tarimo ypatumai, bet ir gramatinės žodžių sudarymo, kaitos ir vartosenos taisyklės: linksniavimas, asmenavimas, priešdėlių, priesagų, prielinksnių vartosena ir kita.

Šiame tyrime buvo bandyta atsižvelgti į priešdėlius ir šiek tiek – į galūnes (žr. komentarus punkte Galūnės). Galūnės yra sudėtingesnis reiškinys, nes jų įvairovė yra gerokai didesnė negu priešdėlių, ir jų pradžia nesutampa su skiemens pradžia.

Priešdėliai. Paėmę dažniau pasitaikančius priešdėlius (at-, ant-, ap-, api-, apy-, be-, į-, iš-, ne-, nebe-, nu-, nuo-, pa-, po-, par-, per-, pra-, pro-, pri-, prie-, prieš-, su-, są-, te-, tebe-, už-) bei dažnesnes jų kombinacijas, taip pat ir su dalelyte „si“, pažymime visus žodžius, kurių pradžia sutampa su bent viena iš sudarytų kombinacijų ir turi be jų bent vieną papildomą skiemenį (balsių grupę). Tokiu būdu sudarytą žymę galima interpretuoti kaip „potencialaus priešdėlio“ (trumpai

PP) savybę. Tarp žodžių su *PP* savybe, žinoma, pasitaiko ir žodžiai, neturintys priešdėlio. Skiemenavimui svarbūs tie atvejai, kai *PP* savybės interpretacija keičia žodžio skiemenavimą: jeigu laikoma, kad priešdėlis yra, žodis skiemenuojamas vienaip, o jeigu priešdėlio nėra arba naudojama alternatyvi formaliai galima priešdėlio grupės interpretacija, žodis skiemenuojamas kitaip. Pavyzdžiui, „antrankiai“, „antrokas“, „apatija“, „apuokas“, „prieinantis“, „prieštara“ ir pan. Tačiau žodžių, turinčių savybę *PP*, bet neturinčių (vienareikšmiškai nusakomos) priešdėlio grupės, kuriai esant esminiai keistųsi to žodžio skiemenavimas, santykinis dažnis lietuviškuose tekstuose tarp žodžių su *PP* savybe, tikėtina, yra nedidelis. Šiame darbe šio klausimo netyrėme.

Buvo sukurti minėtų atskirų potencialių priešdėlių indikatoriai (binarieji kintamieji), priešdėlių dalelytės „si“ ir savybės *PP* indikatoriai, taip pat ir potencialios priešdėlio grupės ilgio kintamasis.

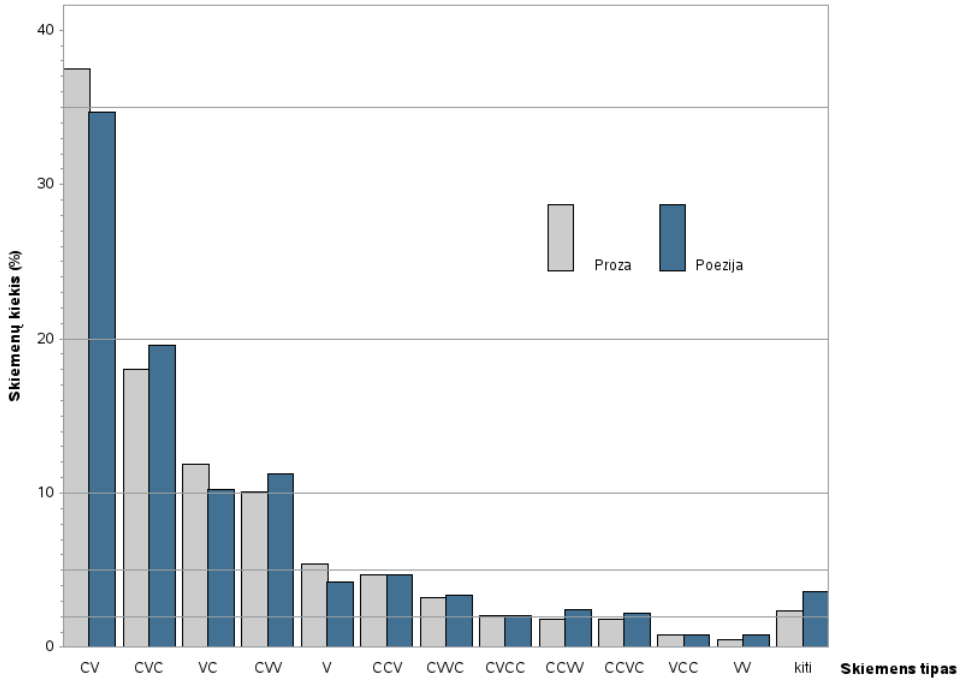
Galūnės. Buvo identifikuotos tekstuose dažniau pasitaikančios galūnės, kurių santykinis dažnis poezijoje ir prozoje pastebimai skyrėsi (remiantis pradinio empirinio tyrimo rezultatais), ir sukurti jų indikatoriai. Rečiau tekstuose pasitaikančios galūnės buvo apjungtos į tam tikras grupes, sukurti tų grupių indikatoriai. Pavyzdžiui: grupė galūnių, kurios naudojamos sutrumpintoje vietininko formoje (-oj,-ėj); analogiška galūnių grupė įnagininkui (-im,-ėm,-om); (galimai) galininko linksnio galūnės (-es); 2-ojo asmens veiksmažodžių sutrumpintos galūnės (-at, -ėt, -it, -ot) ir pan. Bendrai sudaryti 33 binarieji galūnių kintamieji.

Nežiūrint santykinai didelio galūnių kintamųjų skaičiaus ir bandymo kurti interpretuojamus jungtinius galūnių požymius, jie tinkamai neaprašo galūnių vartosenos poezijoje ypatumų lyginant su proza, nes mažai išnaudoja galūnių įvairovę ir gramatinę sandarą.

Skiemens balsių ir priebalsių struktūra. Viena iš svarbių skiemens savybių yra jo balsių ir priebalsių struktūra (sandara). Skiemenį sudaro viena balsių grupė ir pradinė bei galinė priebalsių grupės (tos priebalsių grupės gali būti ir tuščios), sudarytos iš ne daugiau kaip 3 priebalsių, išdėstytų tam tikra tvarka [3, 9, 11]. Kai kurie specifiniai žodelyčiai, pavyzdžiui, „tssss“, „šššššš“, „hmm“ ir pan., gali neturėti ir balsių grupės. Skiemens balsių ir priebalsių (trumpai *CV*) struktūrą įprasta koduoti raidėmis „C“ (*Consonant* – priebalsis) ir „V“ (*Vowel* – balsis) [6, 11]. Pavyzdžiui, žodžių „gra-žus“, „a-kies“, „skris-ti“ skiemenų *CV* struktūros kodai atitinkamai yra CCV-CVC, V-CVVC, CCCVC-CV. Buvo sukurti dažniau pasitaikančių skiemenų *CV* struktūrų (CV, CVC, VC, CVV, V, CCV, CVVC, CVCC, CCVV, CCVC, VCC, VV) indikatoriniai kintamieji (žr. 1 pav.).

Skiemens balsių grupės. Skiemens skambesį lemia jo pagrindą sudaranti balsių grupė (balsis ar dvibalsis). Buvo išskirtos tokios skiemens balsių grupių savybės:

1) *minkštumas* – požymis nusako, ar prieš balsį ar dvibalsį yra minkštumo ženklas (požymis *soft*),
2) *skambumas* – požymis rodo, ar balsių grupė turi savyje balses „a“, „o“ ar „ė“,
3) *ilgosios* – ilgųjų raidžių „y“, „ū“ požymis,
4) *nosinės* – nosinių raidžių požymis,



1 pav. Dažniau pasitaikančių skiemens *CV* struktūrų proporcijų poezijoje ir prozoje palyginimas

5) *dvibalsiai* – požymis rodo, ar balsių grupė yra dvibalsis (šis požymis perdengia skambumo požymį).

Žodžio charakteristikos. Sudarant klasifikatorių buvo naudojamos ir kai kurios bendros žodžio charakteristikos, susijusios su skiemenavimu. Kelios jų jau buvo minėtos: *PP* savybės indikatorius, potencialaus priešdėlio grupės ilgis, tam tikrų žodžio galūnių indikatoriai.

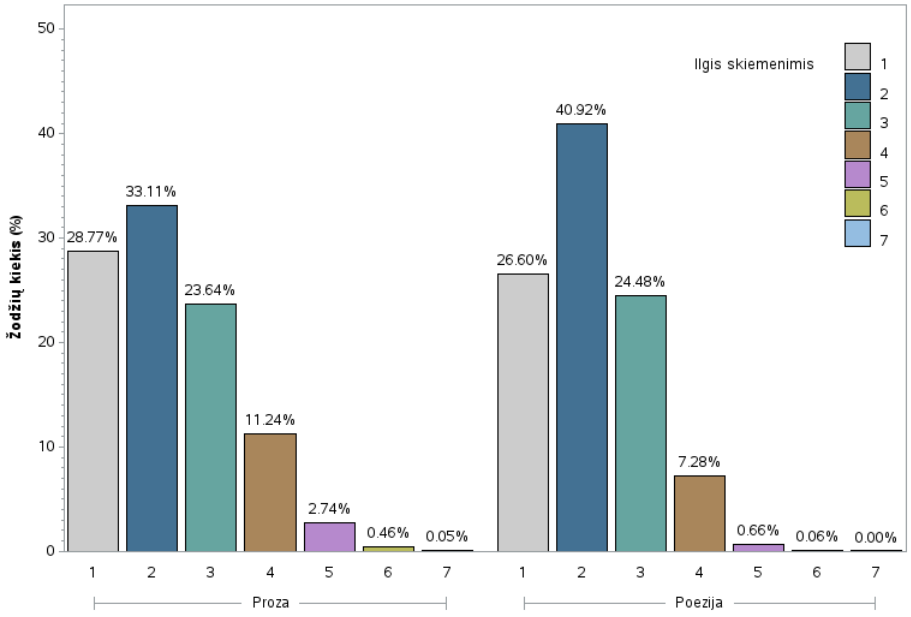
Be to, buvo naudotos tokios žodžio ilgio charakteristikos: žodžio ilgis raidėmis ir jo (dvejetainis) logaritmas, žodžio ilgis skiemenimis (žr. 2 pav.).

Indikatorius *stopw* identifikuoja žodžius, kurie priklauso lietuvių kalbos *tarnybinių žodžių* (*stopwords*) rinkiniui.[3] Tai bendri labai dažnai pasitaikantys atitinkamos kalbos žodžiai, betarpiškai nesietini su teksto turiniu.

Skiemens vieta. Girdenio ir Karosienės [6, 4] atlikti tyrimai rodo, kad skiemenų vartosena priklauso nuo skiemens vietos žodyje. Kaip ir minėtieji autoriai, atskyrėm pirmuosius, paskutiniuosius ir tarpinius žodžių skiemenis. Atskirą skiemenų grupę sudarė vienskiemeniai žodžiai. Apibrėžiant pirmuosius ir tarpinius skiemenis buvo atsižvelgta į *PP* savybę: pirmuoju skiemeniu tokiuose žodžiuose laikomas pirmas skiemuo po potencialaus priešdėlio.

Konkretūs skiemenys. Buvo sukurti dažniau pasitaikančių konkrečių skiemenų, kurių pasitaikymo dažniai (remiantis atliktu pradiniu empiriniu tyrimu) paste-

[3] Lietuviškų tarnybinių žodžių rinkinys yra paimtas iš https://github.com/quanteda/stopwords.



2 pav. Žodžių su skirtingu skiemenų skaičiumi proporcijų poezijoje ir prozoje palyginimas

bimai skiriasi poezijos ir prozos tekstuose, indikatoriai, iš viso 115 indikatorių. Nemažą tų skiemenų dalį sudaro vienskiemeniai tarnybiniai žodžiai. Tik nedaugelis iš jų pasirodė informatyvūs konstruojant klasifikatorių, atskiriantį poeziją nuo prozos.

5 Klasifikatoriaus apmokymas

Klasifikatoriaus apmokymas vykdomas dviem etapais.

Pirmajame etape agreguotiems apmokymo aibės duomenims pritaikius logistinę regresiją su *paieška pirmyn* (*forward selection*) [1] įvertinama skiemens priklausymo poezijai tikimybė $Ppoez$, atitinkamas tiesinis prediktorius η, jo standartinis nuokrypis $std(\eta)$ ir kitos susijusios statistikos ([21], procedūra PROC LOGISTIC).

Antrajame etape kiekvienam apmokymo aibės fragmentui suskaičiuojamos kelios pirmajame etape gautų įverčių $Ppoez$, η, $std(\eta)$ statistikos (vidurkis, mediana, kvartiliai ir kt.) ir jų pagrindu, vėlgi naudojant logistinę regresiją, bet su automatiniu statistiškai nereikšmingų kintamųjų pašalinimu (*backward selection* – modelio kintamųjų atrinkimas atbuline tvarka, [1]), sudaromas fragmento klasifikavimo į klases „poezija“ ir „proza“ požymis $Cpoez$.

Agregavimas. Kadangi planuojama sukurti daug pagalbinių kintamųjų, tai svarbus uždavinys yra sumažinti duomenų skaitmeninę apimtį neprarandant aktualios informacijos. Remiantis tuo, kad klasifikatoriuje bus naudojama tik atskirų žodžių informacija nepriklausomai nuo jų vietos tekste, natūralus būdas sumažinti apmokymui

naudojamo duomenų rinkinio dydį yra pradinius duomenis agreguoti pagal žodžius: imami visi skirtingi apmokymo imties žodžiai ir suskaičiuojami jų dažniai atskirai poezijos tekstuose ir atskirai prozoje. Tokiu būdu agreguotas duomenų rinkinys turi apie 750 tūkst. eilučių vietoje virš 2.5 mln. eilučių pradiniame rinkinyje.

Pirmasis apmokymo etapas. Logistinės regresijos modelis binariojo kintamojo *poezija* reikšmėms prognozuoti buvo parinktas taikant *paieškos pirmyn* procedūrą su maždaug 250 potencialių modelio kintamųjų. Kintamųjų sąveikos (*interactions*) į modelį nebuvo įtrauktos, tačiau iš pagrindinių žodžio ir skiemens kokybinių požymių – savybės *stopw* ir *PP*, skiemens vieta žodyje, skiemens balsių grupės tipas ir minkštumas (savybė *soft*) – buvo sudarytos nesikertančios skiemenų klasės *sClass*. Apjungus neskaitlingas minėtų pagrindinių požymių kombinacijas, skiemenų klasės kintamasis *sClass* reprezentavo 83 klases. Konstruojant klasifikatorių logistinės regresijos modelio statistiškai reikšmingi kintamieji buvo parenkami ir modelio parametrai įvertinami kiekvienai skiemenų klasei atskirai. Pasitaikė klasių, kuriose nė vienas kintamasis nebuvo statistiškai reikšmingas, o kai kuriose klasėse į modelį buvo įtraukta virš 50 kintamųjų.

Dvi svarbios klasifikatorių savybės, nusakančios jų tinkamumą ir efektyvumą yra jų *jautrumas* (*sensitivity*) ir *specifiškumas* (*specifity*) [1].

Apibrėžimas. Klasifikatorių vadinsime *subalansuotuoju*, jeigu klasifikatoriaus *jautrumas* yra (apytiksliai) lygus jo *specifiškumui*.

Subalansuotasis klasifikatorius maksimizuoja klasifikavimo naudingumo funkciją $\min(sensitivity, specifity)$, jautrumo ir specifiškumo reikšmių minimumą. Ji praktiškai beveik nepriklauso nuo klasifikuojamų populiacijų proporcijos mokymo ar testavimo imtyse, jeigu tik jose abiejų populiacijos elementų yra pakankamai daug.

Daugumoje klasių parinkto *subalansuotojo klasifikatoriaus* teisingo spėjimo tikimybė, įvertinta apytiksliu *kryžminio patikrinimo* metodu ([21], procedūra PROC LOGISTIC), svyravo tarp 0.6 ir 0.7, tačiau kai kuriais atvejais įgijo kraštines reikšmes 0.55 ir 0.8.

Įvertintos skiemenų teksto žodžiuose tikimybės priklausyti poezijai *Ppoez*, atitinkamos *tiesinio prediktoriaus* reikšmės η bei jo standartinio nuokrypio $std(\eta)$ įverčiai buvo išsaugoti ir prijungti prie pradinio tekstų rinkinio.

Antrasis apmokymo etapas. Pirmojo etapo rezultatas – pradinis duomenų rinkinys, suskaidytas į 100 žodžių fragmentus, kuriuose prie kiekvieno žodžio, kuris pasitaikė apmokymo imtyje, ir jo skiemenų yra pridėtos įvertintos tikimybės *Ppoez* ir kitos apmokymo metu suskaičiuotos statistikos (η, $std(\eta)$). Paprastas būdas jau pirmajame etape gautos apmokymo informacijos pagrindu atlikti fragmentų klasifikavimą yra, pavyzdžiui, apskaičiuoti įvertintų tikimybių *Ppoez* vidurkį (arba medianą) kiekviename fragmente ir, pasirinkus tinkamą kritinę reikšmę, fragmentą priskirti poezijai, jeigu tas vidurkis (arba mediana) pasirinktą kritinę reikšmę viršija, priešingu atveju – jį priskirti prozai. Tokiu būdu pavyksta gauti *subalansuotąjį klasifikatorių*, kurio įvertinta klaidos tikimybė gerokai mažesnė už 10%, tačiau tokiu būdu sudaryto klasifikatoriaus jautrumas ir specifiškumas testiniams duomenims yra labai jautrus klasifikatoriaus kritinės reikšmės pokyčiams.

Antrasis apmokymo etapas, faktiškai klasifikavimo taisyklės *stabilizavimas*, buvo vykdomas tokiu būdu. Remtasi tik pirmajame etape gautais skiemenų tikėtinumų ly-

ginant poeziją su proza įverčiais *Ppoez*, η, $std(\eta)$. Pagrindinis kintamasis buvo įvertintos tikimybės *Ppoez* vidurkio $\bar{p}$ fragmente *logit* transformacija $pLgt := \log(\bar{p}/(1-\bar{p}))$. Taip pat buvo naudojami kintamieji: *Ppoez* ir η standartiniai nuokrypiai bei tiesinio prediktoriaus η pozicinės statistikos fragmentuose (kvartiliai *eta_Q1*, *eta_Q2*, *eta_Q3* bei 5%, 10%, 90% ir 95% procentiliai *eta_P5*, ..., *eta_P95*).

Fragmentų duomenims pritaikyta logistinė regresija su atsako kintamuoju *poezija* ir minėtais aiškinančiaisiais kintamaisiais, kurie buvo įtraukti į modelį naudojant *atbulinę* kintamųjų atrinkimo procedūrą su kritine tikimybės reikšme 0.001, paliko tik 2 kintamuosius – *pLgt* (natūralu!) ir *eta_P95*. 2 lentelėje pateikta informacija apie minėtų kintamųjų atitinkamų parametrų didžiausio tikėtinumo įvertinius. Atsižvelgus į kintamojo *eta_P95* įvertinto koeficiento neigiamą ženklą jo įtraukimas į modelį skirtas sumažinti retų „labai poetiškų“ skiemenų „teigiamą“ įtaką (žr. 2 lentelę).

2 lentelė. Antrojo etapo logistinės regresijos modelis: didžiausiojo tikėtinumo įvertiniai.

Parametras	Įvertis	Standartinė paklaida	Wald χ^2 statistika	p-reikšmė
Laisvasis narys	104.50	6.91	228.93	<.0001
pLgt	44.12	3.22	188.07	<.0001
eta_P95	−6.33	1.15	30.10	<.0001

Remiantis antrajame etape sudarytu logistinės regresijos modeliu buvo suskaičiuotos mokymo imties fragmentų tikimybių, kad fragmentas priklauso poezijos kūriniui, prognozės $\hat{p}$. Tų prognozių pagrindu buvo sudarytas *subalansuotasis klasifikatorius Cpoez*:

$$Cpoez = \begin{cases} 1, & \text{kai } \hat{p} \geqslant c, \\ -1, & \text{kai } \hat{p} < c, \end{cases}$$

kurio teisingo klasifikavimo apmokymo imtyje tikimybė, įvertinta apytiksliu kryžminio patikrinimo metodu ([21], procedūra PROC LOGISTIC), siekė 97%, ir jos naudingumo funkcija $\min(sensitivity, specifity)$ intervale (0.066, 0.084) mažai priklausė nuo c parinkimo. Optimali kritinė reikšmė $c = 0.079$.

6 Testavimo rezultatų apžvalga

Parinktos klasifikavimo taisyklės *Cpoez* su $c = 0.079$ padarytų klaidų statistika fragmentams iš *apmokymo imties* ($use = 1$), *bendrosios testavimo aibės* ($use = -1$) ir *originaliosios testavimo aibės* ($use = -2$) yra pateikta 3 lentelėje.

3 lentelė. Klasifikatoriaus *Cpoez* su kritine reikšme $c = 0.079$ padarytų klaidų statistika.

Klaidos tipas	Apmokymo aibė	Bendroji testavavimo aibė	Originalioji testavimo aibė
$poezija = 1 \neq Cpoez$	2.06% (7 iš 340)	3.07% (5 iš 163)	8.06% (10 iš 124)
$poezija = -1 \neq Cpoez$	2.69% (153 iš 5678)	4.53% (128 iš 2828)	2.47% (96 iš 3880)

Iš 3 lentelės matyti, kad klasifikatorius *Cpoez* su kritine reikšme $c = 0.079$ originaliojoje testavimo aibėje pastebimai daugiau fragmentų, negu galima būtų tikėtis

subalansuoto klasifikatoriaus atveju, priskyrė prozai (jautrumas 91.94%, o specifiškumas 97.55%), tuo tarpu bendroje testavimo aibėje atvirkščiai – šiek tiek daugiau fragmentų, negu galima būtų tikėtis, buvo priskirta poezijai (jautrumas 96.93%, specifiškumas 95.47%). Matyt, šį poslinkį sąlygojo gana mažas ir nehomogeniškas pagal dydį tyrime naudotų poezijos kūrinių rinkinys. Tačiau pagal padarytų klasifikavimo klaidų bendrą kiekį klasifikatoriaus *Cpoez* veikimo statistika originaliojoje testavimo aibėje net geresnė (klasifikavimo klaidos sudaro 2.65%) negu bendrojoje (klasifikavimo klaidos sudaro 4.45%). Taigi, galima teigti, kad klasifikatorius *Cpoez* su kritine reikšme $c = 0.079$ bendrojoje ir originaliojoje testavimo aibėje duoda palyginamus rezultatus. Tai dera su darbe keliama hipoteze, kad skiemenų statistika daug mažiau susijusi su kūrinių turiniu ir kitomis autorių ar kūrinių specifinėmis ypatybėmis.

Įdomu, kad visi 10 fragmentų iš originaliosios testavimo aibės, klaidingai priskirti prozai, priklauso Erlicko eilėraščiams (iš viso 108 fragmentai). Neteisingai priskirti prozai 5 bendrosios testavimo aibės fragmentai: 3 (iš 24) Erlicko eilėraščių fragmentai ir 2 (iš 44) fragmentai iš Maironio eilėraščių rinkinio „Pavasario balsai“.

Panaudojant testavimo aibės duomenis galima bandyti klasifikatoriui *Cpoez* parinkti geresnę kritinę reikšmę. 4 lentelėje yra pateikti *Cpoez* veikimo rezultatai, kai kritinė reikšmė $c = 0.056$. Matome, kad klasifikavimo taisyklė su šia kritine reikšme yra geriau subalansuota. Bet ir šiuo atveju visi poezijos fragmentai (7 iš originaliosios testavimo aibės ir 4 iš bendrosios testavimo aibės), neteisingai priskirti prozai, priklauso Erlicko eilėraščiams.

4 lentelė. Klasifikatoriaus *Cpoez* su kritine reikšme $c = 0.056$ padarytų klaidų statistika.

Klaidos tipas	Apmokymo aibė	Bendroji testavimo aibė	Originalioji testavimo aibė
$poezija = 1 \neq Cpoez$	2.06% (7 iš 340)	2.45% (4 iš 163)	5.65% (7 iš 124)
$poezija = -1 \neq Cpoez$	3.21% (182 iš 5678)	5.83% (165 iš 2828)	3.30% (128 iš 3880)

Sudarytą naują subalansuotą testiniams duomenims klasifikavimo taisyklę pritaikėme kūrinių, kurie nebuvo priskirti nei poezijai, nei prozai, t.y. fragmentų su požymiu $use = 0$, klasifikavimui. Klasifikatorius „nusprendė“, kad net 6.81% jokiam žanrui nepriskirtų fragmentų yra priskirtini poezijai. Tarp jų: 100% Salomėjos Neries poemos „Eglė žalčių karalienė“, apie 52% Adomo Mickevičiaus poemos „Gražina“ ir apie 81% poemos „Konradas Valenrodas“, apie 92% Balio Sruogos sakmės „Giesmė apie Gediminą“, apie 60% Kosto Ostrausko mikrodramos „Jūratė ir Kastytis“, apie 32% Kristijono Donelaičio poemos „Metai“, beveik 96% Sigito Poškaus poezijos eksperimentų knygos „Nebaigta pasaka“, apie 19% Antano Vaičiulaičio literatūrinės pasakos „Nidos žvėrys“, beveik 65% Marcelijaus Martinaičio pjesės „Pelenų antelė“, 55% Viljamo Šekspyro tragedijos „Romeo ir Džuljeta“.

Detaliau interpretuoti ir komentuoti skiemenų savybes, kurios įtakoja jų santykinai dažnesnį ar retesnį vartojimą poezijos kūriniuose, keblu ir, matyt, nelabai prasminga dėl tų savybių sudėtingų tarpusavio sąryšių. Bet galima įvardyti tokius bendrus gana hipotetinius pastebėjimus:

(1) Skiriant poeziją nuo prozos daug informatyvesni yra *vidiniai žodžių skiemenys*.
(2) Poezijoje rečiau vartojami ilgi skiemenys ir žodžiai.
(3) Poezijoje rečiau vartojami atviri 2-jų raidžių skiemenys ir dažniau – uždari 3-jų raidžių skiemenys.

(4) Poezijoje dažniau vartojami vienskiemeniai tarnybiniai žodžiai.
(5) Poezijoje rečiau vartojami žodžiai su potencialiu priešdėliu (*PP* savybe).
(6) Poezijoje dažniau vartojamos sutrumpintos, „nukąstos“ galūnės, ir rečiau – galūnės, susijusios su įnagininko ir vietininko linksniais.
(7) Poezijoje rečiau vartojami įvardžiai (vienskiemeniai ar priklausantys tarnybinių žodžių sąrašui).
(8) Sudarinėjant logistinės regresijos modelį buvo statistiškai reikšmingi kai kurių konkrečių skiemenų indikatoriai, pavyzdžiui, skiemens „žvaigž“, kuris aiškiai identifikuoja ir patį žodį. Tokie atvejai rodo, kad sudarytas klasifikatorius *Cpoez* išsaugo tam tikrą informaciją apie kūrinio turinį. Šią, atsižvelgiant į tyrimo tikslus, nepageidautiną klasifikatoriaus savybę gana paprasta pataisyti tiesiog neįtraukiant skiemens „žvaigž“ indikatoriaus į modelio kintamųjų sąrašą. Bet *praktiškai* ši problema turėtų būti sprendžiama žymiai praplečiant tekstyną, ypač poezijos kūriniais, galbūt taip pat ir susiaurinant laikotarpį, kurį reprezentuoja tekstyno kūriniai, o svarbiausia – tobulinant klasifikatoriaus konstravimo metodiką.

7 Išvados

Pavyko sudaryti klasifikatorių, kuris yra apmokytas naudojant vien tik atitinkamo teksto skiemenų ar su skiemenavimu susijusiomis žodžių savybėmis ir klasifikuodamas 100 žodžių teksto fragmentus į klases „poezija“ ar „proza“ daro mažiau kaip 5% klaidų. Tai, kad klasifikatorius pakankamai gerai veikė ir su originaliaisiais testavimo fragmentais, ir su testavimo fragmentais tų kūrinių, kurių dalis fragmentų buvo naudoti apmokymui, dera su darbe keliama hipoteze, kad skiemenų statistika daug mažiau susijusi su kūrinių turiniu ir kitais autorių ar kūrinių savitumais. Tai atspindi ir klasifikatoriaus kai kurie reikšmingi kintamieji, kurie rodo, kad vidurinieji žodžių skiemenys yra informatyvesni skiriant poeziją nuo prozos negu pirmieji. Pastarieji yra labiau susiję su žodžio prasme, nes daugumos žodžių šaknys yra vienskiemenės arba dviskiemenės.

Verta paminėti ir kai kuriuos kitus, tikėtina, labiau poezijai nei prozai būdingus požymius: poezijoje rečiau vartojami ilgi skiemenys ir žodžiai, taip pat žodžiai su potencialiu priešdėliu, atviri dviejų raidžių skiemenys bei galūnės, susijusios su įnagininko ir vietininko linksniais, ir dažniau vartojami uždari trijų raidžių skiemenys, vienskiemeniai tarnybiniai žodžiai, sutrumpintos galūnės.

Būtina pabrėžti, kad tyrime naudotas poezijos kūrinių rinkinys yra gana mažas ir nehomogeniškas kūrinių apimties prasme. Todėl šis darbas yra tik skiemenų statistikos galimybių studija ir iliustracija. Sprendžiant poezijos ir prozos atskyrimo uždavinį naudojant skiemenų statistiką reiktų išsamesnio tyrimo, kuris remtųsi praplėstu ir geriau subalansuotu kūrinių rinkiniu bei patobulintu ir labiau automatizuotu klasifikatoriaus konstravimu ir apmokymu.

Padėka

Autoriai dėkoja dr. Karolinai Kanišauskienei už patarimus ir pagalbą ruošiant šį darbą spaudai.

Literatūra

[1] A. Agresti. *Categorical Data Analysis*. Wiley & Sons, 2002. ISBN 9780471249689. https://doi.org/10.1002/0471249688.

[2] D. Daukantaitė, G. Raškinis. Sprendimo medžių panaudojimas skiemenavimo problemai spręsti. In *Informacinės technologijos kalbų inžinerijoje, konferencijos pranešimų medžiaga*, pp. 53–57. Kauno technologijos universitetas, 2006.

[3] A. Girdenis. *Teoriniai lietuvių fonologijos pagrindai*. Mokslo ir enciklopedijų leidybos institutas, Vilnius, 2003.

[4] A. Girdenis, V. Karosienė. Skiemens ir žodžių pirmųjų ir paskutinių fonemų dažnumas bendrinėje lietuvių kalboje. *Baltistica*, **39**(2):213–231, 2004. https://doi.org/10.15388/NA.2019.5.5.

[5] J. Kapočiūtė-Dzikienė, A. Utka, Šarkutė. Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. *Kalbotyra*, **66**, 2014.

[6] V. Karosienė, A. Girdenis. Skiemens ir žodžių pirmųjų ir paskutinių fonemų dažnumas bendrinėje lietuvių kalboje. *Baltistica*, **36**(2):253–266, 2001. https://www.lituanistika.lt/content/34853.

[7] P. Kasparaitis. Automatic stressing of the Lithuanian nouns and adjectives on the basis of rules. *Informatica*, **12**(2):315–336, 2001. https://doi.org/10.3233/INF-2001-12210.

[8] P. Kasparaitis. *Lietuvių kalbos kompiuterinė sintezė*. Daktaro disertacija, Vilniaus universitetas, 2001. http://www.mif.vu.lt/~pijus/publikacijos/KaspDis.pdf.

[9] P. Kasparaitis. *Kompiuterinė lingvistika. Skiemenavimas ir žodžių kėlimas*. Vilniaus universitetas, 2005. https://www.mif.vu.lt/~pijus/CL/Skiemen.pdf.

[10] A. Kazlauskienė. Intervokaliniai priebalsiai: vienanarės ir dvinarės grupės. *Kalbų studijos. Technologija*, **11**:36–42, 2007. https://www.lituanistika.lt/content/17494.

[11] A. Kazlauskienė, G. Raškinis. Lietuvių kalbos fonologinio skiemens struktūrinių modelių dažnumas. *Žmogus ir žodis: mokslo darbai. Didaktinė lingvistika*, **10**(1):24–31, 2008. https://www.lituanistika.lt/content/17113.

[12] A. Kazlauskienė, G. Raškinis, A. Vaičiūnas. *Automatinis lietuvių kalbos žodžių skiemenavimas, kirčiavimas, transkribavimas*. Alka, Kaunas, 2010. ISBN 978-9955-12-630-0. https://www.lituanistika.lt/content/29362.

[13] M. Kroutikov. Tex hyphenation pattern generator, 2016. https://github.com/pgmmpk/pypatgen.

[14] R. Kubiak. Implementing patgen in python, 2019. http://www.gust.org.pl/bachotex/2019-pl/presentations/rkubiak-1-2019.pdf.

[15] M. Lapėnaitė-Gedvilė, K. Piaseckienė, M. Radavišius. Tekstą nehomogeniškumo tyrimas naudojant žymeklius. *Liet. statistikos darbai*, **54**:92–100, 2015. https://doi.org/10.15388/LJS.2015.13884.

[16] F.M. Liang. *Word Hy-phen-a-tion by Com-pu-ter*. Phd dissertation, Stanford University Department of Computer Science, STAN-CS-83-977, 1983. https://www.tug.org/docs/liang/liang-thesis.pdf.

[17] R. Merkytė. Skiemenų ir fonemų skaičiaus lietuvių kalbos žodžiuose savitarpio priklausomybės tyrimas. In *Eksperimentinė ir praktinė fonetika*, pp. 73–84. Vilnius, 1974.

[18] R. Merkytė, V. Kalinka. Apie V. Fukso lingvistinių elementų susidarymo dėsnį (rusų kalba). *Liet. matem. rink.*, **8**(2):279–287, 1968.

[19] G. Raškinis, A. Kazlauskienė. Automatinis skiemenavimas: problemos ir jų sprendimas. *Kalbų studijos*, **15**:71–76, 2009. https://hdl.handle.net/20.500.12259/57452.

[20] P. Sojka, P. Ševeček. Hyphenation in tex — quo vadis? *TUGboat*, **16**(3):280–289, 1995. https://www.fi.muni.cz/usr/sojka/papers/tug95b.pdf.

[21] SAS Institute Inc. 2013. *SAS/STAT ® 13.1 User's Guide*. SAS Institute Inc., Cary, NC, 2013. https://dokumen.tips/documents/sasstat-131-users-guide-this-document-is-an.

[22] E. Stamatatos, N. Fakotakis, G. Kokkinakis. Text genre detection using common word frequencies. In *COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics*, pp. 808–814. Association for Computational Linguistics, Stroudsburg, 2000. https://aclanthology.org/C00-2117.pdf.

[23] V.V. Statulevičius, Y. Haralambous. Lithuanian hyphenation patterns, Vilnius, March 4 1992. https://ltex.lt/apie-mus/.

[24] A. Utka. Labai dažnų lietuvių kalbos žodžių ir žodžių formų ypatybės. *Lituanistica*, **61**(1):48–55, 2005. https://doi.org/https://www.lituanistika.lt/content/46.

SUMMARY

Discriminating poetry and prose using syllable statistics

G. Murauskas, M. Radavičius

The aim of the paper is to construct a universal classifier to discriminate short Lithuanian text excerpts of poetry from that of prose. Here the universality means that the classifier is relatively insensitive to a text content and author's style. Since syllables represent phonetic properties and are less sensitive to text content as compared to words, the classifier training is based on frequencies of syllables in texts to be classified. The text data is taken from digitized library http://ebiblioteka.mkp.emokykla.lt. The error rate of the trained classifier applied to testing excerpts of 100 words is less than 5%.

Keywords: logistic regression; automatic syllabification; cross-validation; training; classification error