Seimo posėdžių stenogramų tekstynas autorystės nustatymo bei autoriaus profilio sudarymo tyrimams

Jurgita Kapočiūtė-Dzikienė
Informatikos fakultetas
Vytauto Didžiojo universitetas
Vileikos g. 8-511
LT-44404 Kaunas, Lietuva
El. paštas: jurgita.k.dz@gmail.com

Andrius Utka
Kompiuterinės lingvistikos centras
Vytauto Didžiojo universitetas
K. Donelaičio g. 52-206
LT-44244 Kaunas, Lietuva
El. paštas: a.utka@hmf.vdu.lt

Ligita Šarkutė
Viešosios politikos ir administravimo institutas
Kauno technologijos universitetas
K. Donelaičio g. 20-217
LT-44239 Kaunas, Lietuva
El. paštas: ligita.sarkute@ktu.lt

Anotacija

Straipsnyje pristatome Seimo posėdžių stenogramų tekstyną, parengtą specialiu formatu, tinkančiu įvairiems autorystės nustatymo tyrimams. Tekstyną sudaro apie 111 tūkstančių tekstų (24 milijonai žodžių), kurių kiekvienas atitinka vieną parlamentaro pasisakymą eilinės sesijos posėdžio metu bei apima 7 Lietuvos Respublikos Seimo kadencijas: nuo 1990 metų kovo 10 dienos iki 2013 metų gruodžio 23 dienos. Pasisakymų tekstai sugrupuoti pagal autorius į 147 grupes, todėl tinka individualių autorių autorystės nustatymo tyrimams; jie suskirstyti pagal autorių amžiaus grupes, lytį ar politines pažiūras, todėl tinka autorių profilio sudarymo tyrimams. Trumpas tekstas neatskleidžia jo autoriaus kalbėjimo stiliaus, yra daugiaprasmiškas kitų autorių atžvilgiu, todėl į tekstyną įtraukti ne trumpesni nei 100 žodžių tekstai. Kiekvieną autorių atitinkantis tekstų rinkinys turi būti išsamus ir reprezentatyvus, todėl įtraukti autoriai, pasisakę ne mažiau kaip 200 kartų. Visi tekstai automatiškai lemuoti, morfologiškai bei sintaksiškai anotuoti, suskaidyti simbolių n-gramomis, surinkta statistinė informacija. Straipsnyje pademonstruota, kaip sukurtas tekstynas gali būti panaudotas individualių autorių autorystės nustatymo bei autorių profilio sudarymo tyrimams, naudojant prižiūrimo mašininio mokymo metodus. Tekstyno struktūra taip pat leidžia taikyti neprižiūrimo

mašininio mokymo metodus, patogi taisyklinių-loginių metodų kūrimui bei įvairioms lingvistinėms analizėms.

Raktažodžiai: Seimo posėdžių stenogramos, autorystės nustatymo tekstynas, stilometrija, individualių autorių autorystės nustatymas, autorių profilio nustatymas

1 Įvadas

Kiekvieno žmogaus rašymo stilius (šablonai, naudojami sakinių formavimui, žodyno turtingumas, frazeologizmai, gramatinės ar sintaksinės klaidos) yra savotiškas jo „piršto antspaudas“. Rašymo stiliumi pradėta domėtis jau 1439 metais (Renna 2014), o jį tyrinėjantis stilometrijos mokslas padarė ypač didelį šuolį pastaraisiais dešimtmečiais. Šį progresą paskatino pats tokių tyrimų poreikis, kurį daugiausia lėmė elektroninių tekstų, ypač anoniminių, atsiradimas.

Vieni stilometrijos uždaviniai sprendžia konkretaus autoriaus autorystės nustatymo problemas: pavyzdžiui, teismo lingvistai nagrinėja, kas internetiniame forume atskleidė konfidencialią įmonės informaciją; kas atsiuntė grasinančio turinio elektroninį laišką, kurio adresas visiškai neinformatyvus; ar kompiuteryje rastą atsisveikinimo laišką iš tiesų parašė pats savižudis; kuris iš socialiniame tinkle prisistatančių asmenų iš tiesų yra užsimaskavęs pedofilas. Kiti stilometrijos uždaviniai apsiriboja autoriaus profilio sudarymu, t. y. autoriaus charakteristikų, tokių kaip amžius, lytis, socialinis statusas, gimtoji kalba, emocinė būsena ir kita, nustatymu. Pavyzdžiui, rinkodaros specialistai siekia išsiaiškinti, kokio amžiaus ar lyties vartotojai labiausiai domisi jų produkcija, iš kriminalinio teksto ištrauktos charakteristikos apie jo autorių policijos pareigūnams padeda sudaryti detalesnį įtariamojo portretą.

Jeigu autorystės nustatymas apsiribotų tik autoriaus verifikacijos tyrimais (Koppel ir Schler 2004, 63), kai turint anoniminį tekstą reikia nustatyti, ar jį parašė mums gerai pažįstamas autorius, ar ne, jis būtų lengvai įveikiamas žmogui. Uždavinys tampa gerokai sudėtingesnis, kai turimi keli šimtai ar net tūkstančiai galimų autorių (autorių-kandidatų): net ir labai reprezentatyvi kiekvieno iš jų rašytų tekstų imtis vargiai padeda nustatyti naujo nežinomo teksto autorystę. Žmogui tiesiog per sunku apdoroti tokius milžiniškus informacijos kiekius, atpažinti kiekvieno iš autorių kalbėjimo stilių ypatumus bei surasti skirtumus tarp jų. Čia į pagalbą ateina automatiniai tyrimo metodai.

2 Autorystės nustatymo tyrimų apžvalga

Visus automatinius autorystės nustatymo metodus galima suskirstyti į taisyklinius-loginius (angl. *rule-based*) ir statistinius – t. y. mašininio mokymo (angl. *machine learning*)

metodus. Taisyklinių-loginių metodų atveju naudojamas žodynas bei rankiniu būdu žmogaus-eksperto sukonstruoti lingvistinių taisyklių šablonai, kurių dėka ir nustatoma tekstų autorystė. Mašininio mokymo metodai gali būti prižiūrimi (angl. *supervised*) ir neprižiūrimi (angl. *unsupervised*). Prižiūrimi mašininio mokymo metodai iš anksto apibrėžtoms tekstų grupėms (grupe gali būti konkretaus autoriaus tekstai, konkretaus amžiaus grupės autorių tekstai ir pan.) geba automatiškai sukonstruoti skiriančiąsias taisykles. Neprižiūrimo mašininio mokymo metodai skiriančiąsias taisykles konstruoja neturėdami informacijos apie tekstų grupes, todėl jų siūlomas „sugrupavimas“ nebūtinai sutampa su žmogaus skirstymu.

1964 metais paskelbta Mosteller ir Wallace (1964) Federalisto užrašų (pagrindinio JAV Konstitucijos interpretavimo šaltinio) studija, pasiūliusi alternatyvą tuo metu populiariems taisykliniams-loginiams metodams, padarė perversmą autorystės nustatymo tyrimų srityje. Pasiūlyta nauja paradigma – automatiniai prižiūrimo mašininio mokymo metodai (Kotsiantis 2007), skirti tekstų klasifikavimui (Sebastiani 2002), jie pradėti taikyti autorystės nustatymo bei autoriaus profilio sudarymo tyrimams (Stamatatos 2009). Šių metodų populiarumą lėmė tai, jog jie efektyviai veikia turint dideles autorių imtis, kai žmogui-ekspertui tiesiog per sunku surasti skirtumus tarp tiek daug grupių ir sukonstruoti kiekvieną iš jų apibūdinančius taisyklių šablonus. Norint taikyti prižiūrimo mašininio mokymo metodus, tereikia turėti mokymo imtį: tekstinių dokumentų aibę $D=\{d_1, d_2, .., d_n\}$ ir apibrėžtą grupių (vadinamų klasėmis) rinkinį $C=\{c_1, c_2, ..., c_m\}$, kur kiekvienas iš dokumentų priskirtas tik vienai jį atitinkančiai klasei: $d \in c$. Individualių autorių autorystės nustatymo atveju klasių turime tiek, kiek turime autorių-kandidatų. Autoriaus profilio sudarymo atveju klasifikavimo uždavinį reikia spręsti kiekvienai iš autorių charakteristikų atskirai, pavyzdžiui, lyties iš teksto nustatymo uždavinio atveju turime dvi klases: vyrišką bei moterišką; amžiaus nustatymo atveju turime tiek klasių, kiek yra amžiaus grupių ir t. t. Vėliau kiekvienas tekstinis dokumentas d automatiniu būdu paverčiamas rinkiniu $x = \{x_1, x_2, \ldots, x_N\}$, kurio elementai x_i apibūdina skirtingas teksto savybes. Pavyzdžiui, jei savybėmis x_i laikysime visus D tekstuose naudojamus žodžius, turėsime žodžių rinkinį (angl. *bag-of-words*), įprastai su įsimintomis tų žodžių pasikartojimų reikšmėmis kiekviename iš d. Pavyzdžiui, x_1 elementas atitinka žodį „puikus“, o x_2 – „geras“, tuomet $x = \{x_1, x_2, ... \} = \{2, 1, ...\}$ reikš, jog žodis „puikus“ šiame dokumente pasikartoja du kartus, o žodis „geras“– vieną. Savybių rinkinys x su priskirta klase c, sudaro mokymo pavyzdį: $<x, c>$. Automatiniai prižiūrimo mašininio mokymo metodai mokymo pavyzdžius automatiškai apibendrina ir transformuoja į taisyklių rinkinį (vadinamą modeliu M), nusakantį, kaip efektyviai atskirti vienas klases nuo kitų: $M(|x|) \rightarrow c$. Sukurtas modelis vėliau gali būti taikomas nežinomų tekstų klasės (konkretaus autoriaus ar autoriaus profilio charakteristikos) nustatymui. Populiariausi automatinių prižiūrimo mašininio mokymo metodų pavyzdžiai: paprastasis daugianaris

Bejesas (angl. *Naive Bayes Multinomial*) (Lewis ir Gale 1994), atraminių vektorių mašina (angl. *Support Vector Machine*) (Cortes ir Vapnik 1995), k-artimiausių kaimynų metodas (angl. *k-Nearest Neighbors*) (Cover ir Hart 1967) ir kt.

Į naudojamų savybių rinkinį gali būti įtraukiami ne tik *D* leksikono žodžiai, bet įvairi kita leksinė, simbolių, morfologinė, sintaksinė bei semantinė informacija. Pačiuose pirmuose autorystės nustatymo tyrimuose naudotos labai primityvios statistinės metrikos, tokios kaip žodžio ar sakinio ilgis (Mendenhall 1887) ir (Yule 1938); žodžių konkrečiame tekstiniame dokumente ir visų skirtingų autoriaus leksikono žodžių santykis (De Vel ir kt. 2001), tačiau jei neįtraukiama jokia kita informacija, šios metrikos dažniausiai nėra patikimos.

Nors aptartasis žodžių rinkinys yra viena efektyviausių anglų kalbos teksto savybių, ypač klasifikuojant tekstus pagal temas arba reiškiamus sentimentus (Pang, Lee, Vaithyanathan 2002), vis dėlto autorystės nustatymo tyrimams tai nėra geriausias pasirinkimas. Įsivaizduokite, jei autoriai-kandidatai savo tekstuose naudoja dar ir skirtingą tematiką: tuomet užuot sprendžiant autorystės nustatymo uždavinį, iš tiesų yra sprendžiamas klasifikavimo į temas uždavinys; todėl, kad žodžių rinkinys būtų patikimas savybių rinkinys, jį reikia naudoti ypač atsakingai, pavyzdžiui, prieš tai pašalinti su aptariamomis temomis susijusius žodžius.

Vietoje visų žodžių kai kuriuose tyrimuose naudojami tik funkciniai žodžiai (tokie kaip prielinksniai, įvardžiai, jungtukai, dalelytės, jaustukai ar artikeliai) (Argamon ir Levitan 2005), kurie nepriklauso nuo aptariamos temos. Skirtingi mokslininkai naudoja skirtingus funkcinių žodžių sąrašus (nuo 150 žodžių (Abbasi ir Chen 2005, 71) iki 675 žodžių (Argamon ir kt. 2007, 815)) dažnai pateikdami mažai informacijos, kuo remiantis tie sąrašai buvo sudaryti. Nors funkciniai žodžiai ir nepriklausomi nuo temos, tačiau kai kurie jų yra labiau būdingi konkretiems funkciniams stiliams (Utka 2006): į šį faktą būtina atsižvelgti, jei skirtingi autoriai naudoja skirtingus funkcinius stilius, kad užuot sprendžiant autorystės nustatymo uždavinį nebūtų sprendžiamas funkcinių stilių klasifikavimo uždavinys.

Jei autorių tematika ir yra panaši, žodžių rinkinys įprastai nėra patogus savybių rinkinys dėl didelio kiekio skirtingų naudojamų to paties žodžio formų, tai ypač aktualu stipriai kaitomoms kalboms, tokioms kaip lietuvių kalba, kurioje žodžio formos nusakomos skirtingomis galūnėmis. Žodžių rinkiniai tampa milžiniški (kiekviena forma tai skirtingas x_i), o sukurti modeliai – neefektyvūs. Leksikono „mažinimo“ problema sprendžiama naudojant lemavimo (žodžiai paverčiami pagrindine forma) (Gamon 2004, 612) ar galūnių nukarpymo (vietoj žodžių paliekami tik jų kamienai) įrankius (Sanderson ir Guenter 2006).

Žodžių (lemų, kamienų) rinkinys talpina informaciją tik apie pavienius elementus, todėl visiškai ignoruojamas jų kontekstas bei žodžių išsidėstymo tvarka sakinyje. Įprastas sprendimo būdas – vietoj žodžių naudoti žodžių n-gramas (žodžių po n rinkinius, kai $n > 1$). Pavyzdžiui, žodžių rinkinio atveju ($n = 1$) frazė „autorystės nustatymo tyrimai“ būtų transformuota į x_0 = „autorystės“, x_1 = „nustatymo“, x_2 = „tyrimai“; žodžių bigramų ($n = 2$) atveju į junginius po du žodžius: x_0 = „autorystės nustatymo“, x_1 = „nustatymo tyrimai“; žodžių trigramų ($n = 3$) atveju į junginius po tris žodžius: x_0 = „autorystės nustatymo tyrimai“. Kalboms, turinčioms griežtą sakinio struktūrą, žodžių n-gramų naudojimas nepagerina autorystės nustatymo tyrimų rezultatų (Coyotl-Morales ir kt. 2006, 849); tačiau kalboms, turinčioms santykinai laisvą sakinio struktūrą (tokioms kaip lietuvių kalba), ši savybė yra rekomenduotina. Žodžių išsidėstymo tvarka gali padėti atskleisti teksto autorių: skirtingi autoriai turi savo mėgstamas konstrukcijas, pavyzdžiui: „aš manyčiau“ ar „mano manymu“, „aš manau“ ar tiesiog „manau“ ir t. t.

Vietoj (ar šalia) leksinių savybių dažnai naudojamos simbolių savybės. Primityvios apsiriboja didžiųjų / mažųjų raidžių kiekių skaičiavimu, skyrybos ženklų skaičiavimu (Zheng ir kt. 2006), tačiau populiariausios – simbolių n-gramos. Pavyzdžiui, turint $n = 5$ frazė „trys autoriai“ būtų sudalyta į elementus po penkis simbolius (įskaitant ir tarpo simbolį, žymimą „_“): „trys_“, „rys_a“, „ys_au“, „s_aut“, „_auto“, „autor“, „utori“, „toria“, „oriai“. Simbolių n-gramos šiuo metu yra pati svarbiausia autoriaus rašymo stilių apibūdinanti savybė: ji nepriklauso nuo kalbos (teisingai parinktas n leidžia atsikratyti kaitomomis galūnėmis), todėl nereikalauja išorinių gramatinių įrankių (skirtų lemavimui ar galūnių nukarpymui); gali būti naudojama kalboms, neturinčioms tarpų tarp žodžių (tokioms kaip kinų ar japonų); atspari klaidoms (pavyzdžiui, „autorystė“ ir klaidingai užrašytas žodis „autoristė“ vis tiek turės labai daug bendrų simbolių n-gramų); leidžia atsižvelgti į kontekstą tarp greta einančių žodžių. Ne veltui simbolių n-gramos savo tikslumu dažnai aplenkia net leksines savybes (Luyckx 2010). Kuri reikšmė n leistų generuoti didžiausią tikslumą iš anksto nėra žinoma, todėl jos nustatymas reikalauja atskiro eksperimentinio tyrimo. Eksperimentiškai nustatyta, jog simbolių tetragramos yra pats geriausias pasirinkimas lietuvių kalbai, tačiau tyrimai atlikti tekstinių dokumentų klasifikavimo į temas uždaviniui (Kapočiūtė-Dzikienė ir kt. 2012, 1403).

Morfologinės (Zhao ir Zobel 2007) bei sintaksinės savybės (Stamatatos ir kt. 2000, 480) dar labiau leidžia pagerinti autorystės nustatymo ar autoriaus profilio sudarymo rezultatus, tačiau tik naudojant jas kaip priedą prie leksinės ar simbolių informacijos. Automatinio autorystės nustatymo atveju siekiama automatizuoti visas jos grandis, todėl užuot rankiniu būdu morfologiškai ar sintaksiškai anotuojant tekstus, naudojami automatiniai anotavimo įrankiai, o bendras tikslumas priklauso ir nuo šių įrankių veikimo tikslumo. Morfologinių bei sintaksinių įrankių veikimo tikslumas – ne vienintelė problema. Ne visi šie įrankiai yra laisvai prieinami, o tai smarkiai apriboja galimų tyrimų įvairovę.

Autorystės nustatymo ar autoriaus profilio sudarymo tyrimams parengti tekstynai dažnai naudojami kaip etalonas naujo siūlomo metodo efektyvumui įvertinti. Etaloninių tekstynų pavyzdžiai: šiuolaikinės graikų kalbos tekstynas autorystės nustatymo tyrimams (Stamatatos ir kt. 2000); olandų kalbos tekstynas autoriaus asmenybės tipo nustatymo tyrimams (Luyckx ir Daelemans 2008) ir kt. Rezultatai, gauti su etaloniniais tekstynais, yra lengvai palyginami tarpusavyje, todėl tokie tekstynai dažnai naudojami mokslininkų varžybų, tokių kaip AAC (angl. *Ad-hoc Authorship Attribution Competition*) (organizuota 2004, 2006 metais) ar PAN (angl. *Uncovering Plagiarism, Authorship and Social Software Misuse*) (organizuojama nuo 2007 iki dabar), metu, kuriose siekiama pasiūlyti tiksliausiai veikiantį metodą ir kartu išbandyti, kokios metodikos yra tinkamiausios skirtingoms kalboms, jų tipams, žanrui. Kaip matome, tinkamai parengtų tekstynų (kurie taip pat galėtų būti etaloniniai) nauda automatiniams autorystės nustatymo ar autoriaus profilio sudarymo tyrimams yra neabejotina.

Lietuvoje autorystės nustatymo tyrimų pradžia galima laikyti 1971 metus, kai pirmą kartą aptarta autoriaus individualaus stiliaus (idiolekto) sąvoka (Pikčilingis 1971). Per tą laiką pasiūlyta daug įvairių tyrimo metodikų. Pavyzdžiui, teismo lingvistikos tekstų analizei palengvinti siūloma atsižvelgti į rašybos, skyrybos klaidas, tarmybes (Dambrauskaitė 1972), naudojamą sinonimiką (Žalkauskienė 2000) ir kt.; elektroninių laiškų autorystės nustatymams būtina atskirti temos raišką ir autoriaus įpročius, elektroniniams laiškams būdingą ir nebūdingą raišką, atsižvelgti į autoriaus kalbinę raišką žyminčius žodžius, grafikos ženklus ir kt. (Žalkauskaitė 2012, 160). Ankstesni lietuviškų tekstų autorystės nustatymo tyrimai daugiausia apsiriboja lingvistinėmis analizėmis, kai turima nedidelė autorių imtis, todėl jiems visiškai nebūtinas specialiai parengtas tekstynas. Vis dėlto tam, kad pasiūlyti lingvistiniai metodai galėtų būti automatizuoti, o ateityje, galbūt, komercializuoti, prieš tai būtina įvertinti jų tikslumą. Galima atlikti ir žmogiškąjį ekspertinį vertinimą, tačiau paprastai tokie vertinimai turi didelių trūkumų: užima daug laiko, dažniausia yra sunku išvengti subjektyvumo, sunku juos pakartoti. Todėl efektyviausias metodų vertinimas yra gautų rezultatų lyginimas su etaloniniu tekstynu. Tokį tekstyną galima naudoti tiek lyginamosioms analizėms, tiek ir naujų siūlomų metodų tobulinimui.

Šio straipsnio tikslas: pristatyti pirmąjį autorystės nustatymo ir autoriaus profilio sudarymo tyrimams parengtą norminės lietuvių kalbos tekstyną bei įvertinti tekstyno naudojimo galimybes su prižiūrimo mašininio mokymo metodais. Rengiant tekstyną siekta, kad jis būtų patogus tiek lingvistinėms analizėms, tiek jį naudoti tiriant įvairius automatinius metodus. Parengtas tekstynas yra prieinamas Vytauto Didžiojo universiteto Kompiuterinės lingvistikos centro puslapyje[1].

[1] http://tekstynas.vdu.lt/page.xhtml?id=projects-current

3 Seimo posėdžių stenogramų tekstynas

Seimo posėdžių stenogramų tekstynas rengtas 7-ių Lietuvos Respublikos Seimo kadencijų (nuo 1990 metų kovo 10 dienos iki 2013 metų gruodžio 23 dienos) pagrindu. Parengtą tekstyną sudaro apie 111 tūkstančių norminės lietuvių kalbos tekstų (iš viso apie 24 milijonus žodžių), kurių kiekvienas atitinka vieną parlamentaro pasisakymą eilinės sesijos posėdžio metu.

Trumpas tekstas nėra pakankamai informatyvus, kad būtų galima efektyviai nustatyti jo autorystę, todėl vieni mokslininkai siūlo naudoti 2500 žodžių (Eder 2013, 2), kiti 500 žodžių (Koppel ir kt. 2007, 1265); treti pasiekia neblogus rezultatus (viršijančius atsitiktinę ribą bei didžiausios klasės tikimybę) ir su labai trumpais, kurių vidurkis vos 60 žodžių, tekstais (Luyckx 2011, 2; Mikros ir Perifanos 2011, 1). Kadangi vieningo susitarimo tarp mokslininkų dėl teksto ilgio nėra, be to, nenorėjome stipriai palengvinti sprendžiamo autorystės nustatymo uždavinio, todėl į tekstyną įtraukėme ir trumpus, tačiau ne trumpesnius nei 100 žodžių tekstus.

Tekstynas parengtas, kad tiktų:

A. Individualių autorių autorystės nustatymo tyrimams (žr. 1 lentelę). Tekstai sugrupuoti į 147 grupes: kiekviena grupė atitinka konkretų autorių, o tai grupei priskirti tekstai – to autoriaus tekstai. Tekstyno reprezentatyvumui užtikrinti įtraukėme tik tuos autorius, kurie pasisakė ne mažiau nei 200 kartų.

B. Autoriaus profilio sudarymo tyrimams pagal šias charakteristikas:

    a) amžiaus grupę (žr. 2 lentelę). Tekstai sugrupuoti pagal 6 amžiaus grupes. Amžiaus grupių skaičių bei patį suskirstymą įprastai lemia turimas uždavinys. Pasirinkome grupavimą, dažniausiai naudojamą socialiniuose tyrimuose. Atitinkamą grupavimą pagal amžių naudoja didžiausi socialinių mokslų duomenų archyvai Europoje, pvz., Vokietijos GESIS – Leibnico socialinių mokslų institutas. Analogiškos amžiaus grupės naudojamos ir Lietuvos humanitarinių ir socialinių mokslų duomenų archyve (LiDA). Rankiniu būdu surinkta informacija apie autorių gimimo datas leido suskirstyti turimus tekstus pagal amžiaus grupes. Skirstydami tekstus nepašalinome ribinių atvejų (t. y. patenkančių į intervalą ± 1 ar 2 metai aplink amžiaus grupės ribą), kaip kartais yra daroma (Rangel ir Rosso 2013, 7), kad nepalengvintume sprendžiamo uždavinio.
    b) lytį (žr. 3 lentelę). Tekstai sugrupuoti į 2 grupes.

c) politines pažiūras (žr. 4 lentelę). Tekstai suskirstyti į 3 grupes, rankiniu būdu surinkus informaciją apie parlamentarų politines pažiūras bei tų pažiūrų pasikeitimus. Politinių pažiūrų grupės išskirtos pagal tradicinę politinės kairės-dešinės ideologijos skalę. Šiame spektre politinės ideologijos atrodo taip: komunizmas-socializmas-liberalizmas-konservatizmas-fašizmas (Heywood 1992, 9). Parlamentarų grupavimui pagal politines pažiūras naudotos tik tradicinės ideologijos neįtraukiant kraštutinių – politinės Seimo narių pažiūros pagal priklausymą konkrečioms partijoms priskirtos socialistinėms arba socialdemokratinėms, t. y. kairiosioms; liberaliosioms, t. y. centro; konservatyviosioms, t. y. dešiniosioms. Jei parlamentaras nepriklausė jokiai politinei partijai, jo politinės pažiūros ekspertiniu būdu buvo priskirtos vienai iš grupių pagal iš jo pasisakymų ir ankstesnės veiklos numanomas pažiūras.

Autorius	**Tekstų kiekis**	**Žodžių kiekis**	**Skirtingų žodžių kiekis**	**Vid. teksto ilgis žodžiais**
JURŠĖNAS Č.	7 779	1 454 762	58 078	187,0
KUBILIUS A.	5 093	1 076 084	57 728	211,3
ANDRIUKAITIS V. P.	3 932	926 247	58 673	235,6
VESELKA J.	3 018	672 644	55 597	222,9
VIDŽIŪNAS A.	2 654	455 756	29 612	171,7
STEPONAVIČIUS G.	2 539	456 598	31 239	179,8
DEGUTIENĖ I.	2 491	484 498	32 220	194,5
PEČELIŪNAS S.	2 430	458 652	37 773	188,7
MUNTIANAS V.	2 241	384 398	24 633	171,5
SYSAS A.	2 130	451 444	36 621	211,9
SAKALAS A.	2 024	400 853	35 719	198,0
DAGYS R. J.	1 934	406 769	37 530	210,3
PAULAUSKAS A.	1 808	361 556	35 138	200,0
LANDSBERGIS V.	1 732	595 325	57 239	343,7
RAZMA J.	1 719	316 059	33 980	183,9
...	...	...	...	...
VISI	**110 908**	**23 908 302**	**314 653**	**215,6**

1 lentelė. Seimo posėdžių stenogramų tekstynas individualių autorių autorystės nustatymo tyrimams[2]

[2] Siekiant skaitytojo neperkrauti per dideliu informacijos kiekiu, lentelėje pateikiama tik 15 iš 147 daugiausiai pasisakiusių autorių.

Amžiaus grupė	Tekstų kiekis	Žodžių kiekis	Skirtingų žodžių kiekis	Vid. teksto ilgis žodžiais
Iki 29	707	145 682	21 438	206,1
30–39	16 549	3 577 765	117 504	216,2
40–49	28 700	6 139 128	162 174	213,9
50–59	42 325	9 199 619	209 189	217,4
60–69	17 895	3 867 463	142 708	216,1
Nuo 70	4 732	978 645	65 958	206,8
VISI	**110 908**	**23 908 302**	**314 653**	**215,6**

2 lentelė. Seimo posėdžių stenogramų tekstynas autoriaus profilio pagal amžiaus grupę nustatymo tyrimams

Lytis	Tekstų kiekis	Žodžių kiekis	Skirtingų žodžių kiekis	Vid. teksto ilgis žodžiais
Moteriška	10 727	2 357 596	101 953	219,8
Vyriška	100 181	21 550 706	301 420	215,1
VISI	**110 908**	**23 908 302**	**314 653**	**215,6**

3 lentelė. Seimo posėdžių stenogramų tekstynas autoriaus profilio pagal lytį nustatymo tyrimams

Politinės pažiūros	Tekstų kiekis	Žodžių kiekis	Skirtingų žodžių kiekis	Vid. teksto ilgis žodžiais
Centro	28379	5795499	157552	204,2
Dešiniosios	42233	9362957	210545	221,7
Kairiosios	40296	8749846	196037	217,1
VISI	**110 908**	**23 908 302**	**314 653**	**215,6**

4 lentelė. Seimo posėdžių stenogramų tekstynas autoriaus profilio pagal politines pažiūras nustatymo tyrimams

Visi tekstai parengti specialiu formatu, kad būtų galima tyrinėti, kaip įvairios savybės veikia autorystės nustatymo tyrimų tikslumą. Savybės pateikiamos 5 lentelėje.

Savybės tipas	Savybės	Aprašymas
Pagrindinės stilometrinės charakteristikos	Vidutinis žodžio ilgis dokumente	Simbolių visuose žodžiuose ir žodžių kiekio santykis
	Vidutinis sakinio ilgis dokumente	Žodžių ir sakinių kiekio santykis
	Normalizuotas skirtingų žodžių kiekis	Skirtingų žodžių ir visų žodžių kiekio santykis

Leksinės savybės	Funkciniai žodžiai	Šių kalbos dalių žodžiai: prielinksniai, įvardžiai, jungtukai, dalelytės ir jaustukai. Pastarosios kalbos dalys nustatytos naudojant automatinį morfologinį analizatorių bei lemavimo įrankį Lemuoklis (Zinkevičius 2000) ir (Daudaravičius ir kt. 2007).
	Žodžiai	Visi žodžiai (nepaisant, kokia morfologine forma jie užrašyti).
	Lemos	Žodžiai, paversti pagrindine žodžio forma – lema. Žodžių transformavimui į lemas naudotas Lemuoklis.
Morfologinės savybės	Kalbos dalys	Nurodyta kiekvieno iš žodžių kalbos dalis. Kalbos dalių nustatymui naudotas Lemuoklis.
	Morfologinės žymos	Nurodyta smulkesnė informacija apie kiekvieną iš žodžių: linksnis, giminė, skaičius, laikas ir kt. Šios informacijos nustatymui naudotas Lemuoklis
Sintaksinės savybės	Sintaksinės žymos	Vietoj žodžių nurodytos sintaksinės žymos (naudojant priklausomybių gramatiką): veiksnys, tarinys ir t. t. Žymų nustatymui naudotas surastas efektyvus automatinis metodas (Kapočiūtė-Dzikienė ir kt. 2013), kurio modulis apmokytas su lietuviškuoju medžių banku[3].
Simbolių savybės	Simbolių n-gramos	Tekstinis dokumentas suskaidytas simbolių n-gramomis (prieš tai pašalinus skyrybos ženklus, tačiau palikus tarpo simbolius). Pateikti 6 skirtingi rinkiniai: kai $n = 2, 3, ... 7$.

5 lentelė. Tekstiniuose dokumentuose nurodytos savybės[3]

4 Eksperimentai ir rezultatai

Šiuo straipsniu siekėme ne tik pristatyti naujai parengtą tekstyną, tačiau pademonstruoti, jog jį galima pritaikyti individualių autorių autorystės nustatymo bei autoriaus profilio sudarymo tyrimams. Klasifikavimui pasirinkome vieną populiariausių prižiūrimo mašininio mokymo metodų – paprastąjį daugianarį Bejesą (Lewis ir Gale 1994), kadangi šis metodas:

A. Neatsižvelgia į nesusijusias savybes. Sakykime, kad autoriaus X išskirtinis bruožas –

[3] Medžių bankas sukurtas VDU vykdyto Lietuvos valstybinio mokslo ir studijų fondo Lituanistikos mokslinių tyrimų prioriteto įgyvendinimo 2007–2008 metų programos projekto „Internetiniai ištekliai: anotuotas lietuvių kalbos tekstynas ir anotavimo priemonės (ALKA 2)“ metu.

savybės x_0 naudojimas ($x_0 \neq 0$, kai $x_1 = 0$), o autoriaus Y – savybės x_1 naudojimas ($x_0 = 0$, kai $x_1 \neq 0$). Jeigu likusios savybės X ir Y atveju bus panašios, kuriamam modeliui jos įtakos neturės.

B. Gali veikti, kai turima daug vienodo reikšmingumo savybių. Sakykime, jog kelios savybės yra vienodai reikšmingos sprendžiamam uždaviniui. Metodas nereikalauja nustatyti prioritetų tarp jų, kaip, pavyzdžiui, sprendimų medžių metodo atveju (Quinlan 1986).

C. Labai greitas ir nereikalauja didelių apibendrintų duomenų saugojimo resursų.

D. Įprastai pasirenkamas kaip pradinis metodas (angl. *baseline*) etaloniniam tekstynui įvertinti.

Daugianario paprastojo Bejeso metodo veikimo principas paremtas prielaida, jog nežinomas tekstinis dokumentas d' turi būti priskirtas klasei, su kuria sąlyginė tikimybė $P(c|d')$ yra didžiausia. Sąlyginė tikimybė skaičiuojama:

$$P(c \mid d') = P(c) \prod_{1 \le i \le n_d} P(x_i \mid c)$$

n_d – elementų, esančių d', kiekis;
x_i – i-asis d' elementas.

Pirminė tikimybė $P(c)$ ir sąlyginė tikimybė $P(x_i|c)$ apskaičiuojamos remiantis mokymo imtyje D saugoma informacija: konstruojamas modelis M analizuojant, kaip d, esantys D, yra susiję su skirtingais c.

$P(c)$ apskaičiuojama taip:

$$P(c) = \frac{N_c}{N}$$

N_c – tekstinių dokumentų, esančių D, kurie priklauso c, kiekis;
N – tekstinių dokumentų, esančių D kiekis.

$P(x_i|c)$ apskaičiuojama taip:

$$P(x_i \mid c) = \frac{kiekis(x_i \mid c) + 1}{kiekis(c) + |V|}$$

kiekis $(x_i|c)$ – elemento x_i, priklausančio c iš D, kiekis;
kiekis (c) – elementų, priklausančių c iš D, kiekis;
$|V|$ – skirtingų elementų iš D kiekis (leksikono dydis).

Eksperimentų metu naudojome aprašyto metodo programinę realizaciją, esančią pakete WEKA 3.6 (Hall ir kt. 2009). Rezultatų įvertinimui naudojome 10 dalių kryžminę validaciją (Weiss ir Kulikowski 1991): D esantys pavyzdžiai padalijami į 10 dalių; su 9 dalimis kuriamas modelis, su likusia – įvertinamas jo efektyvumas; eksperimentai kartojami 10 kartų, kad rezultatai būtų gauti vis su nauja testavimo dalimi, vėliau

apskaičiuojamas bendras rezultatas. Rezultatų efektyvumo įvertinimui naudojome tikslumo metriką:

$$Tikslumas = \frac{T_{teisingi}}{T_{visi}}$$

$T_{teisingi}$ – dokumentų d' kiekis, kuriems teisingai nustatyta klasė;
T_{visi} – visų testavimui naudotų dokumentų kiekis.

Eksperimentiškai patikrinome 5-oje lentelėje nurodytus vienos rūšies savybių rinkinius:

a) pagrindines stilometrines charakteristikas;
b) funkcinius žodžius;
c) žodžių n-gramas (nuo $n = 1$ iki $n = 3$);
d) lemų n-gramas (nuo $n = 1$ iki $n = 3$);
e) kalbos dalių n-gramas (nuo $n = 1$ iki $n = 3$);
f) morfologinių žymų n-gramas (nuo $n = 1$ iki $n = 3$);
g) sintaksinių žymų n-gramas (nuo $n = 1$ iki $n = 3$);
h) simbolių n-gramas (nuo $n = 2$ iki $n = 7$)

bei šiuos jungtinius savybių rinkinius:

a) žodžių ir kalbos dalių n-gramas (nuo $n = 1$ iki $n = 3$);
b) lemų ir kalbos dalių n-gramas (nuo $n = 1$ iki $n = 3$);
c) žodžių ir morfologinių žymų n-gramas (nuo $n = 1$ iki $n = 3$);
d) lemų ir morfologinių žymų n-gramas (nuo $n = 1$ iki $n = 3$);
e) žodžių ir sintaksinių žymų n-gramas (nuo $n = 1$ iki $n = 3$);
f) lemų ir sintaksinių žymų n-gramas (nuo $n = 1$ iki $n = 3$).

Sprendėme du uždavinius:

A. Individualių autorių autorystės nustatymo (žr. 1 paveikslą). Eksperimentams naudojome subalansuotą duomenų aibę D: kiekvienam autoriui atsitiktinai atrinkome po vienodą kiekį (t. y. 200) tekstinių dokumentų. Tyrėme tikslumo priklausomybę nuo vis didėjančio autorių kiekio (vis įtraukiant naujus autorius). Eksperimentai buvo atlikti išbandant visas prieš tai nurodytas savybes, tačiau paveiksle pateikiami tik geriausi apibendrinti rezultatai. Pastebėta, jog nepriklausomai nuo autorių kiekio, geriausi rezultatai gauti su leksinėmis savybėmis (lemomis, žodžiais), juos papildžius morfologinėmis žymomis. Metodą galima laikyti efektyviu, kadangi rezultatai viršija didžiausios klasės tikimybę (angl. *majority baseline*), skaičiuojamą:

$\max_c \left(\frac{N_c}{N} \right)$

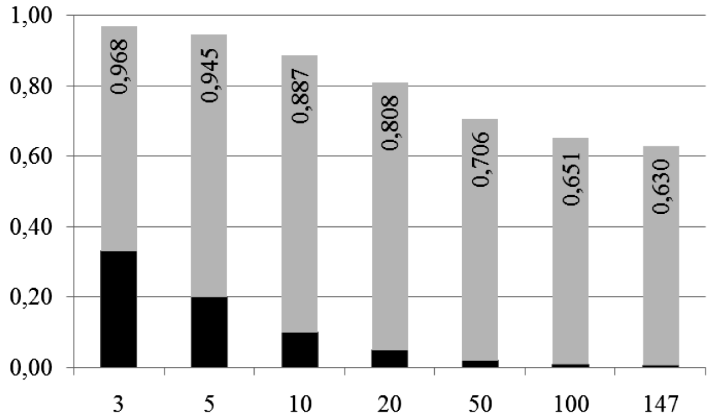
N_c – tekstinių dokumentų, esančių D, kurie priklauso c, kiekis;
N – tekstinių dokumentų, esančių D, kiekis.

ir atsitiktinę ribą (angl. *random baseline*), skaičiuojamą:

$\sum_c \left(\frac{N_c}{N} \right)^2$

N_c – tekstinių dokumentų, esančių D, kurie priklauso c, kiekis;
N – tekstinių dokumentų, esančių D, kiekis.

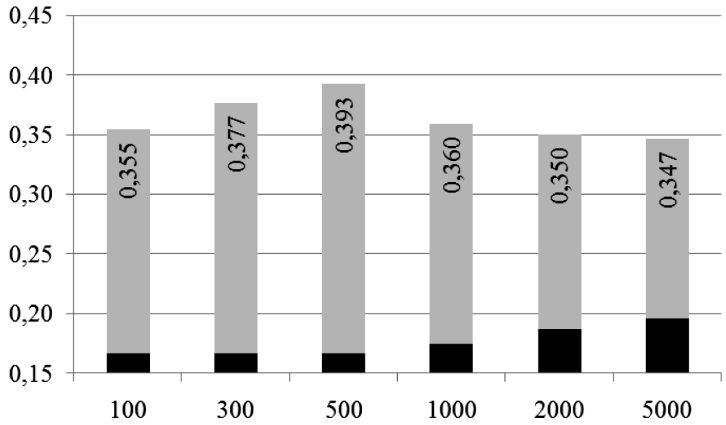
1 paveiksle x ašis nurodo autorių kiekį, y ašis – didžiausią gautą tikslumą. Kiekvieno stulpelio tamsi apatinė dalis žymi maksimalią atsitikinės ribos ar didžiausios klasės tikimybės reikšmę. Didėjant autorių kiekiui tikslumas prastėja (vis sunkiau metodui sukonstruoti skiriančiąsias taisykles): kai turimi 3 autoriai, tikslumas siekia net 96,8%, kai turimi 147 autoriai – tikslumas siekia 63,0%.



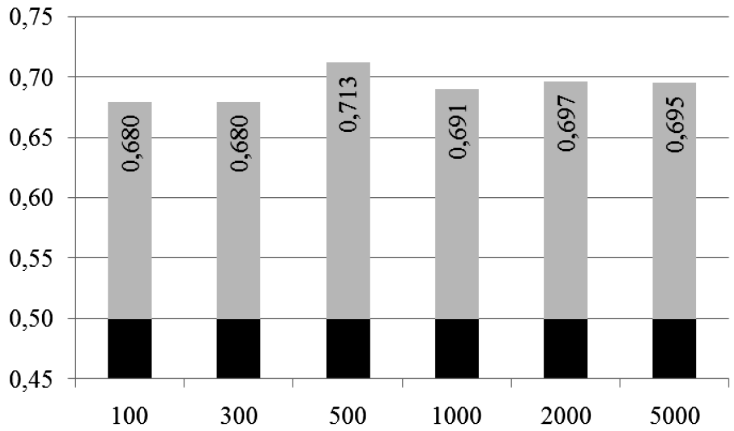
1 paveikslas. Autorystės nustatymo tyrimų rezultatai: tikslumo priklausomybė nuo individualių autorių kiekio

B. Autoriaus profilio sudarymo pagal šias charakteristikas: amžių (žr. 2 paveikslą), lytį (3 paveikslas) ir politines pažiūras (4 paveikslas). Eksperimentams naudojome subalansuotas duomenų aibes. Tyrėme tikslumo priklausomybę nuo duomenų aibės dydžio: atsitiktinai imdami po 100, 300 ir t. t. tekstinių dokumentų kiekvienai iš klasių. Eksperimentai buvo atlikti su visomis prieš tai aprašytomis savybėmis, tačiau paveiksluose pateikiami tik geriausi apibendrinti rezultatai. Geriausi rezultatai su amžiaus charakteristika gauti naudojant žodžius, papildytus kalbos dalimis; geriausi rezultatai su lyties charakteristika gauti naudojant lemas; geriausi rezultatai su politinių pažiūrų charakteristika gauti naudojant lemas arba lemas, papildytas kalbos dalimis. Visi rezultatai viršijo didžiausios klasės tikimybę ir atsitiktinę ribą.

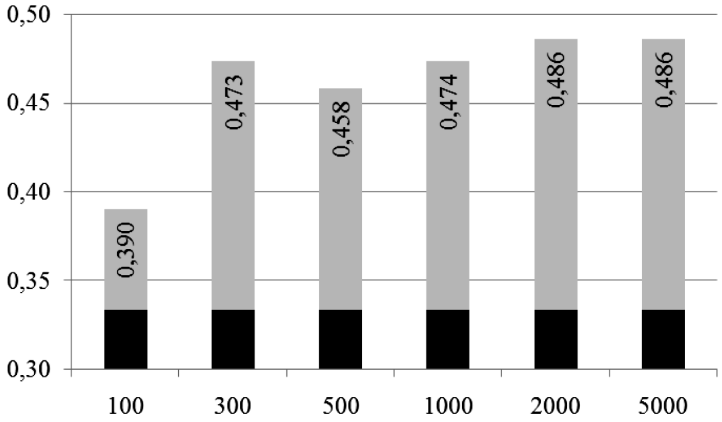
2, 3 ir 4 paveiksluose *x* ašis nurodo naudojamų tekstinių dokumentų kiekį kiekvienoje iš klasių, *y* ašis – didžiausią gautą tikslumą. Nustatant amžiaus grupę ar lytį su paprastuoju Bejesu pakanka mažesnės duomenų aibės: geriausi rezultatai, t. y. 39,3% amžiaus grupės nustatymo atveju, 71,3% lyties nustatymo atveju, gauti, kai klasėse buvo po 500 tekstinių dokumentų. Tuo tarpu geriausi rezultatai, t. y. 48,6% politinių pažiūrų nustatymo atveju, gauti su didele imtimi – 2000 ar 5000 tekstinių dokumentų kiekvienoje iš klasių. Kuo mažiau klasių, tuo lengviau metodui surasti skirtumus tarp klasių, todėl geriausi rezultatai gauti nustatant lytį (2 klasės), o prasčiausi – nustatant amžiaus grupę (6 klasės).



2 paveikslas. Autorių amžiaus grupės nustatymo tyrimų rezultatai: tikslumo priklausomybė nuo imties dydžio



3 paveikslas. Autorių lyties nustatymo tyrimų rezultatai: tikslumo priklausomybė nuo imties dydžio



4 paveikslas. Autorių politinių pažiūrų nustatymo tyrimų rezultatai: tikslumo priklausomybė nuo imties dydžio

5 Apibendrinimai ir išvados

Pristatėme naują specialiu formatu parengtą tekstyną, skirtą individualių autorių autorystės nustatymo bei autoriaus profilio (amžiaus, lyties ir politinių pažiūrų charakteristikoms) sudarymo tyrimams. Tekstynas automatiniu būdu lemuotas, morfologiškai bei sintaksiškai (naudojant priklausomybių gramatiką) anotuotas, tekstai suskaidyti simbolių n-gramomis, apie kiekvieną iš tekstų surinkta statistinė informacija: vidutiniai žodžių bei sakinių ilgiai, normalizuotų skirtingų žodžių kiekiai.

Straipsnyje eksperimentiškai pademonstravome, jog turimas tekstynas gali būti naudojamas jam taikant prižiūrimo mašininio mokymo metodus. Šį tekstyną galima naudoti kaip etaloną kitų metodų (taisyklinių-loginių, neprižiūrimo mašininio mokymo, kitų prižiūrimo mašininio mokymo) tikslumo vertinimui. Tekstynas taip pat tinkamas įvairioms lingvistinėms analizėms.
Tolimesni autorystės nustatymo bei autoriaus profilio sudarymo tyrimai orientuoti dviem kryptimis: 1) į naujų metodų, galinčių pagerinti esamų metodų tikslumą, kūrimą; 2) į naujų tekstynų, ypač orientuotų į šnekamąją kalbą (internetinių dienraščių, socialinių tinklų, interneto forumų ir kt.), kūrimą.

Padėka

Straipsnis parengtas vykdant projektą „Automatiniai autorių ir autorių grupių individualaus stiliaus nustatymo tyrimai (ASTRA)“, kurį finansuoja Lietuvos mokslo taryba (projekto Nr. LIT-8-69).

Duomenų šaltiniai

AAC. *Autorystės nustatymo mokslinių varžybų internetinis puslapis.* Interneto prieiga: http://www.mathcs.duq.edu/~juola/authorship_contest.html.

GESIS. *Leibnico socialinių mokslų instituto puslapis*. Interneto prieiga: http://www.gesis.org.

LiDA. *Lietuvos duomenų archyvas*. Interneto prieiga: http://www.lidata.eu/.

LR. *Lietuvos Respublikos Seimo stenogramų archyvas*. Interneto prieiga: http://www3.lrs.lt/pls/inter/w5_sale.kad_ses.

PAN. *Plagiato, autorystės nustatymo ir autoriaus profilio sudarymo mokslinių varžybų internetinis puslapis*. Interneto prieiga: http://www.uni-weimar.de/medien/webis/research/events/pan-14/.

WEKA 3.6. *Mašininio mokymo metodų programinė realizacija.* Interneto prieiga: http://www.cs.waikato.ac.nz/ml/weka/.

Literatūros sąrašas

Abbasi, Ahmed, Hsinchun Chen. 2005. Applying authorship analysis to extremist-group web forum messages. *IEEE Intelligent Systems* 20 (5), 67–75.

Argamon, Shlomo, Casey Whitelaw, Paul Chase, Sushant Dhawle, Sobhan Raj Hota, Navendu Garg, Shlomo Levitan. 2007. Stylistic text classification using functional lexical features. *Journal of the American Society for Information Science and Technology* 58 (6), 802–822.

Argamon, Shlomo, Shlomo Levitan. 2005. Measuring the usefulness of function words for authorship attribution. *Proceedings of the Joint Conference of the Association for Computers and Humanities and the Association for Literary and Linguistic Computing*.

Cortes, Corinna, Vladimir Vapnik. 1995. Support-vector networks. *Machine Learning* 20, 273–297.

Cover, Thomas, Peter Hart. 1967. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory* 13 (1), 21–7.

Coyotl-Morales, Rosa María, Luis Villaseñor-Pineda, Manuel Montes-y-Gómez, Paolo Rosso. 2006. Authorship attribution using word sequences. *Proceedings of the 11th Iberoamerican Congress on Pattern Recognition*, 844–853.

Dambrauskaitė, Ona Danutė. 1972. *Kriminalističeskoe issledovanie litovskoj pis'mennoj reči. Avtoreferat dissertacij na soiskanie učenoj stepeni kandidata juridičeskih nauk*. Leningrad.

Daudaravičius, Vidas, Erika Rimkutė, Andrius Utka. 2007. Morphological annotation of the Lithuanian corpus. *Proceedings of the Workshop on Balto-Slavonic Natural*

Language Processing: Information Extraction and Enabling Technologies (ACL'07), 94–99.

De Vel, Olivier, Alison Anderson, Malcolm Corney, George Mohay. 2001. Mining e-mail content for author identification forensics. *SIGMOD Record* 30 (4), 55–64.

Eder, Maciej. 2013. Does size matter? Authorship attribution, small samples, big problem. *Literary and Linguistic Computing*, 1–16.

Gamon, Michael. 2004. Linguistic correlates of style: Authorship classification with deep linguistic analysis features. *Proceedings of the 20th International Conference on Computational Linguistics*, 611–617.

Hall, Mark, Eibe Frank, Geoffre Holmes, Bernhard Pfahringer, Peter Reutemann, Ian Witten. 2009. The WEKA data mining software: an update. *SIGKDD Explorations* 11 (1), 10–18.

Heywood, Andrew. 1992. *Political Ideologies*. UK, London: Macmillan.

Kapočiūtė-Dzikienė, Jurgita, Frederik Vaassen, Walter Daelemans, Algis Krupavičius. 2012. Improving topic classification for highly inflective languages. *Proceedings of 24th International Conference on Computational Linguistics* (COLING 2012), 1393–1410.

Kapočiūtė-Dzikienė, Jurgita, Joakim Nivre, Algis Krupavičius. 2013. Lithuanian dependency parsing with rich morphological features. *Empirical Methods in Natural Language Processing – 4th Workshop on Statistical Parsing of Morphologically Rich Languages* (SPMRL'2013), 12–21.

Koppel, Moshe, Jonathan Schler. 2004. Authorship verification as a one-class classification problem. *Proceedings of the Twenty-first International Conference on Machine Learning* (ICML'04), 62–68.

Koppel, Moshe, Jonathon Schler, Elisheva Bonchek-Dokow. 2007. Measuring differentiability: unmasking pseudonymous authors. *Journal of Machine Learning Research* 8, 1261–1276.

Kotsiantis, Sotiris. 2007. Supervised machine learning: a review of classification techniques. *Informatica* 31, 249–268.

Lewis, David, William Gale. 1994. A sequential algorithm for training text classifiers. *Proceedings of Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval* (SIGIR-94), 3–12.

Luyckx, Kim. 2010. *Scalability Issues in Authorship Attribution*. PhD thesis in linguistics. Belgium: University of Antwerp.

Luyckx, Kim. 2011. Authorship attribution of e-mail as a multi-class task (notebook for PAN at CLEF 2011). *Cross-Language Evaluation Forum (Notebook Papers/Labs/ Workshop)*.

Luyckx, Kim, Walter Daelemans. 2008. Personae: a corpus for author and personality prediction from text. *LREC, European Language Resources Association*.

Mendenhall, Thomas Corwin. 1887. The characteristic curves of composition. *Science* 9, 237– 249.

Mikros, George, Kostas Perifanos. 2011. Authorship identification in large email collections: experiments using features that belong to different linguistic levels (notebook for PAN at CLEF 2011). *Cross-Language Evaluation Forum (Notebook Papers/Labs/Workshop)*.

Mosteller, Frederik, David Wallace. 1964. *Inference and Disputed Authorship: the Federalist. Series in Behavioral Science: Quantitative Methods*. USA, Massachusetts: Addison-Wesley.

Pang, Bo, Lillian Lee, Shivakumar Vaithyanathan. 2002. Thumbs up? Sentiment classification using machine learning techniques. *Proceedings of Conference on Empirical Methods in Natural Language Processing* (EMNLP), 79–86.

Pikčilingis, Juozas. 1971. *Kas yra stilius*? Vilnius: Vaga.

Quinlan, Ross. 1986. Induction of decision trees. *Machine Learning* 1 (1): 81–106.

Rangel, Francisco, Paolo Rosso. 2013. Use of language and author profiling: identification of gender and age. *Proceedings of the 10th Workshop on Natural Language Processing and Cognitive Science* (NLPCS-2013).

Renna, Thomas. 2014. Lorenzo Valla and the donation of Constantine in historical context, 1439–40. *Expositions* 8, 11–28.

Sanderson, Conrad, Simon Guenter. 2006. Short text authorship attribution via sequence kernels, Markov chains and author unmasking: an investigation. *Proceedings of the International Conference on Empirical Methods in Natural Language Engineering*, 482–491.

Sebastiani, Fabrizio. 2002. Machine learning in automated text categorization. *ACM Computing Surveys* 34, 1–47.

Stamatatos, Efstathios. 2009. A survey of modern authorship attribution methods. *Journal of the Association for Information Science and Technology* 60 (3), 538–556.

Stamatatos, Efstathios, Nikos Fakotakis, George Kokkinakis. 2000. Automatic text categorization in terms of genre and author. *Computational Linguistics* 26 (4), 471–495.

Utka, Andrius. 2006. Common words as indicators of text functions. *Prace Baltystyczne* 3, 213–224.

Weiss, Sholom, Casimir Kulikowski. 1991. *Computer systems that learn*. San Francisco, California, USA: Morgan Kaufmann.

Yule, George Udny. 1938. On sentence-length as a statistical characteristic of style in prose, with application to two cases of disputed authroship. *Biometrica* 30, 363–390.

Zhao Ying, Justin Zobel. 2007. Searching with style: authorship attribution in classic literature. *Proceedings of the Thirtieth Australasian Computer Science Conference*, 59–68.

Zheng, Rong, Jiexun Li, Hsinchun Chen, Zan Huang. 2006. A framework for authorship identification of online messages: writing style features and classification techniques. *Journal of the American Society of Information Science and Technology* 57 (3), 378–393.

Zinkevičius, Vytautas. 2000. Lemuoklis – morfologinei analizei. *Darbai ir Dienos* 24, 246–273.

Žalkauskaitė, Gintarė. 2012. *Idiolekto požymiai elektroniniuose laiškuose*. Humanitarinių mokslų daktaro disertacija. Vilnius: Vilniaus universitetas.

Žalkauskienė, Anelė. 2000. Lietuviško teksto autoriaus nustatymo metodikos pagrindai. *Jurisprudencija* 18 (10), 113–121.

Corpus of transcribed parliamentary speeches for authorship attribution and author profiling tasks

Jurgita Kapočiūtė-Dzikienė, Andrius Utka, Ligita Šarkutė

Summary

In our paper we present a corpus of transcribed Lithuanian parliamentary speeches. The corpus is prepared in a specific format, appropriate for different authorship identification tasks. The corpus consists of approximately 111 thousand texts (24 million words). Each text matches one parliamentary speech produced during an ordinary session from the period of 7 parliamentary terms starting on March 10, 1990 and ending on December 23, 2013. The texts are grouped into 147 categories corresponding to individual authors, therefore they can be used for authorship attribution tasks; besides, these texts are also grouped according to age, gender and political views, therefore they are also suitable for author profiling tasks. Whereas short texts complicate recognition of author speaking style and are ambiguous in relation to the style of other authors, we incorporated only texts containing not less than 100 words into the corpus. In order to make each category as comprehensive and representative as possible, we included only those authors, who produced speeches at least 200 times. All the texts are lemmatized, morphologically and syntactically annotated, tokenized into the character n-grams. The statistical information of the corpus is also available. We have also demonstrated that the created corpus can be effectively used in authorship attribution and author profiling tasks with supervised machine learning methods. The corpus structure also allows using it with unsupervised machine learning methods and can be used for creation of rule-based methods, as well as in different linguistic analyses.

Įteikta 2014 m. gegužės mėn.