Vilniaus universiteto leidykla

Contents lists available at Vilnius University Press

Information & Media
2023, vol. 97, pp. 25–48

eISSN 2783-6207
DOI: https://doi.org/10.15388/Im.2023.97.57

Teksto analizės įrankio „Voyant Tools“ panaudojimas mokslinės informacijos analizei

Aušra Kairaitytė-Užupė

Vilniaus universitetas, Kauno fakultetas, Lietuva
ausra.kairaityte-uzupe@knf.vu.lt
https://orcid.org/0000-0002-3585-3172

Egidija Ramanauskaitė

Vilniaus universitetas, Kauno fakultetas, Lietuva
egidija.kiskina@knf.vu.lt
https://orcid.org/0000-0001-6814-8667

Vytautas Evaldas Rudžionis

Vilniaus universitetas, Kauno fakultetas, Lietuva
vytautas.rudzionis@knf.vu.lt
https://orcid.org/0009-0007-5683-7546

Santrauka. Straipsnyje pristatomos mokslinės informacijos analizės galimybės taikant kompiuterinę tekstų analizės programą „Voyant Tools“. Nagrinėjamas tekstynas, sudarytas iš 404 „Clarivate Analytics Web of Science“ ir „Scopus ScienseDirect“ duomenų bazėse publikuotų atvirosios prieigos straipsnių, skirtų skaitmeninės humanitarikos problematikai. Straipsnyje aptariami kiekybiniai teksto analizės metodai, atsietojo ir interaktyviojo skaitymo galimybės, kurias suteikia atvirosios prieigos „Voyant Tools“ platformoje integruoti tekstų sisteminimo įrankiai. Straipsnio autoriai pristato problemas, su kuriomis susidūrė atlikdami teksto analizę, taip pat – įvertina analizės rezultatų vizualizavimo naudingumą tyrimui ir interpretacijų paieškai. Kompiuteriniai įrankiai gali pasitarnauti patyrusiems tyrėjams, kurie domisi kiekybiniais teksto analizės metodais, o pradedantiems tyrėjams atsiranda galimybė įgyti pradinių žinių, kurios paskatins giliau domėtis kompiuterine tekstų analize.

Pagrindiniai žodžiai: skaitmeninė humanitarika; kompiuterinė tekstų analizė; atsietasis skaitymas; interaktyvusis skaitymas; „Voyant Tools“.

Scientific Information Analysis Using Text Analysis Tool “Voyant Tools”

Abstract. This article describes the use of “Voyant Tools”, an open access text analysis application, to examine a corpus of articles from open access journals, dealing with the topic of digital humanities. The corpus consisted of 404 articles recorded in the “Clarivate Analytics Web of Science” and “Scopus ScienceDirect” databases. The authors discuss how “Voyant Tools” aids to identify the dominant fields of research through quantitative methods and to reveal the main discourse themes using distant reading and interactive reading capabilities. They also identify some problems encountered during the analyses, and also discuss the usefulness of data

Received: 2022-08-17. **Accepted:** 2023-02-22.

visualization for research and interpretation. Computer tools can be useful for experienced researchers who are interested in quantitative text analysis, as well as for beginners, as it provides an opportunity to acquire basic knowledge that will lead to a deeper interest in textual analysis methods.

Keywords: digital humanities; computational text analysis; distant reading; interactive reading; "Voyant Tools".

Įvadas

Akademinė bendruomenė susiduria su nuolat augančiu mokslinių straipsnių srautu, kurį atveria atvirosios prieigos duomenų bazės. Mokslinės literatūros apžvalgai taikomos įvairios atrankos strategijos ir informacijos sisteminimo metodai[1], įskaitant kompiuterinius metodus. Atrinkus duomenų bazėse publikacijas pagal raktinius žodžius, vis dar lieka gausus jų skaičius, todėl jų skaitymui ir analizei verta pasitelkti kompiuterį.

Šiame straipsnyje nagrinėjama didelės apimties duomenų analizės problema. Norint gerai orientuotis atliekamų tyrimų problematikoje, tyrėjui tenka atsirinkti temą geriausiai reprezentuojančius tekstus ir juos išnagrinėti. Su tokia užduotimi susidūrė ir šio straipsnio autoriai, norėdami apžvelgti „Clarivate Analytics Web of Science" ir „Scopus ScienseDirect" duomenų bazėse esančių humanitarinių mokslų žurnalų straipsnius, atskleidžiančius skaitmeninės humanitarikos veiklas. Skaitmenine humanitarika šiame straipsnyje laikome akademinės veiklos sfer**ą,** susiejan**čią** skaitmenines technologijas ir humanitarinių mokslų disciplinas[2].

Šio straipsnio tikslas – aptarti kompiuterines informacijos apdorojimo galimybes ir rezultatus, panaudojant skaitmeninės humanitarikos veiklas pristatantį tekstyną, kurį sudaro 404 straipsniai.

Tikslui pasiekti užsibrėžėme tokius uždavinius:

1. Taikant raktinius žodžius, iš pasirinktų duomenų bazių atrinkti straipsnius, reprezentuojančius skaitmeninės humanitarikos problematiką, ir sudaryti tekstyną;
2. Apžvelgti tekstyne vyraujančius diskursus taikant kompiuterinę teksto analizę;
3. Įvertinti analizės rezultatų vizualizavimo naudingumą temų ir interpretacijų paieškai.

Tyrime taikyti atsietojo skaitymo[3] (angl. *distant reading*) ir interaktyviojo skaitymo metodai. Atsietasis skaitymas gali būti suprantamas kaip tradicinio teksto (bei kitų duome-

[1] Pvz., J. Bettany-Saltikov (2012) mini aštuonis sisteminės literatūros apžvalgos etapus: (1) temos pasirinkimas, (2) apžvalgos plano sudarymas, (3) objekto patikslinimas ir straipsnių atmetimas, (4) sisteminė literatūros paieška, (5) straipsnių, kurie atsako į pagrindinės temos klausimą, atranka, (6) rezultatų sintezė, apibendrinimas ir pristatymas, (7) diskusija, rezultatų aptarimas ir (8) pasidalinimas jais.

[2] Kadangi skaitmeninės humanitarikos samprata plačiai aptarinėjama mokslinėje literatūroje, tačiau dar nenusistovėjusi, naudinga pasiremti pačių šios akademinės veiklos atstovų rašomu ir redaguojamu šaltiniu – Vikipedija, kurioje pateiktas skaitmeninės humanitarikos ir jos probleminių klausimų apibūdinimas (prieiga per internetą: https://en.wikipedia.org/wiki/Digital_humanities). Interneto svetainėse, skirtose universitetų skaitmeninės humanitarikos veikloms pristatyti, ji dažniausiai siejama su kompiuterinių metodų taikymu socialinių ir humanitarinių mokslų problemoms spręsti ir skaitmeninimu. Pavyzdį žr.: https://www.helsinki.fi/en/digital-humanities.

[3] F. Moretti (2000) pasiūlyta atsietojo skaitymo sąvoka ir idėjos plačiai taikomos skaitmeninės humanitarikos tyrimuose (pvz., taikymo pavyzdžiai išsamiai aptariami Jänicke et al. (2015)), šių idėjų pagrindu įkurta „Stanford Literary Lab" (prieiga per internetą: https://litlab.stanford.edu/), sukurtas įvairias mokslo institucijas vienijantis tinklas „Distant Reading for European Literary History" (prieiga per internetą: https:// www.distant-reading.net).

nų) išreiškimas vizualiomis priemonėmis (grafikais, laikmačiais, žemėlapiais, skalėmis ir kt.) naudojant kompiuterines programas. Atsietasis skaitymas naudingas tyrinėjant didelės apimties literatūrinius, istorinius bei kitus tekstus, nes vizualizuoja teksto ar tekstų masyvo bendruosius bruožus (Jänicke et al., 2015; Keturakis, 2019: 110) bei padeda pažvelgti į jį konceptualiai. Metodo pasirinkimas priklauso nuo tiriamojo klausimo – ką norime sužinoti.

Tyrimui pasitelkta kompiuterinė tekstų analizės programa „Voyant Tools“[4]. „Voyant Tools“ platformoje integruoti kiekybiniai tekstų analizės metodai, todėl tekstas, vizualiai išreikštas skalėmis ir grafikais, tyrėjams gali pasiūlyti naujų ir netikėtų įžvalgų (Van Atteveldt et al., 2019: 162). Pereinant prie interaktyviojo skaitymo, atsiranda galimybė panaudoti skirtingus „Voyant Tools“ platformoje integruotus įrankius, kurie padeda išsamiau peržiūrėti kontekstą.

Straipsnyje pristatysime dalį „Voyant Tools“ platformoje integruotų kiekybinių metodų, supažindinsime su „Voyant Tools“ taikymo tyrimuose galimybėmis; aptarsime mūsų tyrimui naudoto tekstyno sudarymo ir medžiagos sisteminimo priemones ir, pasitelkę statistinę analizę, išnagrinėsime šio tekstyno problematiką; diskusijų dalyje įvardysime aptiktas problemas ir pateiksime išvadas.

Kiekybiniai tekstų analizės metodai

Prieš aptariant „Voyant Tools“ platformą ir jos įrankių taikymo didelės apimties duomenų analizei problematiką, trumpai pristatysime kai kuriuos kiekybinio teksto analizės metodus[5], naudojamus „Voyant Tools“ platformoje, kad mažiau susipažinę skaitytojai galėtų aiškiau suprasti teksto analizės metodų esmę, jų taikymo galimybes ir apribojimus.

Teksto analizės metodai gali būti skirstomi į kokybinius ir kiekybinius. Analizuojant tekstus skaitmeniniu būdu, dėl tekstų prigimties ir daugelio kompiuterinės analizės metodų specifikos dažniau yra taikomi kiekybinės analizės metodai. Kiekybinė teksto analizė suprantama kaip teksto turinio analizė, naudojanti įvairias teksto skaitines charakteristikas. Tokia analizė atliekama kompiuteriu, o šių metodų atsiradimą ir išplėtojimą lėmė didėjančios kompiuterių galimybės ir skaitmeninių tekstų gausa. Esant didesniems duomenų kiekiams galima daryti patikimesnes ir įvairesnes išvadas. Atliekant kiekybinę teksto analizę, keliama prielaida, jog tekstas gali būti charakterizuojamas tam tikrais skaitiniais parametrais, tačiau parametrų interpretacija ir išvados priklauso nuo tyrėjų.

Kiekybinei teksto analizei būdingi keturi etapai: 1) tekstų atranka, vadinama tekstyno sudarymu, 2) tekstų pritaikymas, paverčiant juos kompiuterinėms programoms suprantamu elektroniniu formatu, 3) dokumento apibrėžimas, t. y. nustatomas teksto analizės vienetas,

[4] „Voyant Tools“ (prieiga per internetą: https://voyant-tools.org/) pasirinktas dėl jo funkcionalumo ir atvirosios prieigos. Galima aptikti įvairių įrankių, atliekančių žodžių debesies generavimo funkciją, tokių kaip „WordCloud“ (prieiga per internetą: https://www.wordclouds.com/), „WordCloud Generator“ (prieiga per internetą: https://monkeylearn.com/word-cloud/), „WordArt“ (prieiga per internetą: https://wordart.com/) ir kt. Populiarios sudėtingesnės tekstyno analizės įrankių programos, atliekančios kolokacijų, žodžių dažnių analizę: „AntConc“ (prieiga per internetą: https://www.laurenceanthony.net/software/antconc/), „WordSmith Tools“ (prieiga per internetą: https://lexically.net/wordsmith/), „Sketch Engine“ (prieiga per internetą: https://www.sketchengine.eu/) ir kt.

[5] Plačiau su šių metodų įvairove galima susipažinti, pvz., Jurafsky & Martin (2021).

4) požymių apibrėžimas (identifikuojami žodžiai, lemos, žodžių grupės, žodžių dalys, skirtukai ir t. t.). Toliau pasirinktiems požymiams taikomos kiekybinės arba statistinės procedūros ir interpretuojami gauti rezultatai.

Paprasčiausią kiekybinės analizės metodų grupę sudaro *baziniai deskriptyviniai metodai*. Pirmasis plačiai taikomas analizės metodas – *žodžių dažnių analizė*, kurios metu suskaičiuojama, kiek kartų vienas ar kitas žodis pasikartoja tekste ar tekstuose. Žodžių dažnių analizės metodai gali būti pritaikomi įvairioms kalboms ir skirtingiems tikslams pasiekti, pvz., brukalams identifikuoti (Stringhini et al., 2010), vyraujančioms teksto temoms (Wallach, 2006) ar teksto autoriaus lyčiai nustatyti (Alsmearat et al., 2014).

Pagrindinė žodžių dažnio analizės metodo prielaida – dažniau aptinkami žodžiai yra „svarbesni“, t. y. jie saugo daugiau svarbios teksto informacijos. Prieš atliekant žodžių dažnių analizę, dažnai aptinkami, bet semantiškai mažai reikšmingi žodžiai (pvz., jungtukai, ištiktukai ir pan.) yra pašalinami, o kiti teksto žodžiai suvedami į pagrindines žodžių (pvz., daiktavardžių, būdvardžių, veiksmažodžių) formas[6]. Žodžių dažnių analizė neretai leidžia gana neblogai įvertinti leksinę teksto įvairovę: kuo autoriaus vartojama leksika yra įvairesnė, tuo žodžių dažnių lentelė yra tolygesnė. Tačiau leksinė teksto įvairovė labai priklauso nuo teksto ilgio. Žodžių dažnių analizė yra naudojama įvertinti, kaip lengvai yra skaitomas tekstas. Trumpais sakiniais parašytas tekstas yra skaitomas lengviau, tuo remiantis anglų kalbai empiriškai nustatytas Flesch–Kincaid indeksas vertina skaitomą tekstą (Kincaid et al., 1975):

$$206.835 - 1.015\left(\frac{\text{total words}}{\text{total sentences}}\right) - 84.6\left(\frac{\text{total syllables}}{\text{total words}}\right)$$

Jeigu indekso reikšmė yra 0–30 – tekstą gali skaityti universitetinį išsilavinimą turintys asmenys, 60–70 – visi suaugę asmenys, o esant indeksui nuo 90 iki 100 – tekstą gali skaityti 11 metų vaikai. Panašūs empiriniai indeksai yra sudaryti ir kitoms kalboms[7].

Vienas iš seniausių (pasiūlytų Zelligo Harriso (1954)) ir populiariausių žodžių dažnių analizės metodo adaptavimų yra žodžių krepšio (angl. *bag of words*) algoritmas, naudojamas tekstų panašumui įvertinti arba tekstui priskirti vienai iš galimų kategorijų. Nesuklysime pasakę, kad praktiškai visi kiti žodžių dažnių analizės metodai yra tolesnis šio metodo išplėtojimas. Paprasčiausiu atveju metodas taikomas taip: norint priskirti nežinomą tekstą vienai iš kategorijų, surenkami tekstai, kurių tematikos yra žinomos, ir kiekvienai iš tematikų sudaromas žodžių krepšys, t. y. suskaičiuojama, kaip dažnai tam tikras žodis aptinkamas skirtingos tematikos tekstuose. Tuomet žodžiai surikiuojami pagal dažnius, t. y. sudaromos dažnių lentelės, vadinamos žodžių krepšiais. Toks žodžių krepšys sudaromas ir nežinomam tekstui bei palyginama, į kokios tematikos tekstų žodžių krepšį pastarasis yra panašiausias. Manoma, kad nežinomas tekstas yra tos tematikos, į kurios

[6] Pvz., daiktavardžiai: lapas = lap-o, lap-ui, lap-e; veiksmažodžiai: bėgti = bėg-a, bėg-o; būdvardžiai: gražus = graž-iam, graž-aus.

[7] Pvz., ispanų (Spaulding, 1956), arabų (Daud et al., 2013) ir kt. Žinoma, kad toks indeksas buvo skaičiuotas ir lietuvių kalbai, tačiau viešai publikuoto indekso rasti neteko.

žodžių krepšį yra panašiausias. Pvz., tikimasi, kad sportinės tematikos tekstuose sporto šakų pavadinimai bus aptinkami dažniau negu medicininės tematikos tekstuose.

Vienas iš pagrindinių žodžių krepšio metodo trūkumų yra tai, kad, sprendžiant pasirinktą uždavinį, nėra vertinama santykinė žodžio reikšmė (Jurafsky & Martin, 2021). Pvz., norint identifikuoti teksto tematiką kai kurie žodžiai gali būti dažnai aptinkami įvairių tematikų tekstuose. Jų informatyvumas nėra didelis. Tačiau kai kurie žodžiai dažnai aptinkami tik tam tikros tematikos tekstuose, todėl jų informatyvumas didesnis. Norint įvertinti žodžio svarbą dokumente, naudojamas vadinamasis *tf-idf* koeficientas (angl. *term frequency – inverse document frequency*). Koeficientą pasiūlė Hansas Peteris Luhnas 1957 metais (Luhn, 1957). Koeficientas naudojamas identifikuoti temą (Salton & Buckley, 1988; Zhu et al., 2019), sentimentus (Das & Chakraborty, 2018), finansines klaidas pagal apskaitininkų komentarus (Rudžionis et al., 2022) ir kitiems tikslams.

tf – tai skaičius, rodantis, kiek kartų žodis ar žodžių grupė buvo aptikta tam tikrame dokumente. Natūralu tikėtis, kad kuo dažniau bus vartojamas žodis, tuo žodžio svarba bus didesnė. Tačiau kai kurie žodžiai labai dažnai vartojami visuose dokumentuose ir jų dažnas aptikimas atskirame dokumente mažai informuoja apie žodžio svarbą atskirame dokumente. *idf* dalis įvertina, kaip dažnai žodis aptinkamas visuose dokumentuose. *tf* dalis apskaičiuojama taip:

$$tf = f / \sum f,$$

čia f – kiek kartų tam tikras žodis buvo aptiktas dokumente, o – bendras dokumento žodžių skaičius.

idf dalis apskaičiuojama taip:

$$idf = log(N/d),$$

čia N – visų tekstyne esančių dokumentų skaičius, o d – dokumentų, kuriuose buvo rastas dominantis žodis, skaičius.

tf-idf koeficientas apskaičiuojamas dauginant *tf* ir *idf* dalis:

$$tfidf = tf * idf.$$

Tačiau žodžių dažnių analizės metodai nevertina, kurioje teksto vietoje žodžiai pasitaiko, o atliekant tyrimus, gali būti svarbu žinoti, ar tam tikri žodžiai tekste išsidėsto tolygiai, ar priešingai – koncentruojasi skirtingose teksto dalyse.

Svarbu paminėti, kad humanitarai, besidomintys kompiuterių naudojimu tematinei ar stilistinei tekstų analizei, turi būti giliau negu eilinis vartotojas susipažinę su kompiuterio veikimo logika, su kuria susijusios kompiuterinės teksto analizės galimybės[8]. Kiekybine teksto analize pagrįsti algoritmai yra tinkami specifiniams klausimams atskleisti. Naudingiausia ją naudoti ne izoliuotai, bet derinant su įprastinėmis perspektyvomis. Šiuo atveju kompiuteris gali tapti geru pagalbininku svarstant interpretacijos klausimus.

[8] Plačiau apie kompiuterinės analizės principus žr. „Computer Criticism“ (Smith, 1978: 329–334).

„Voyant Tools“ platforma ir jos įrankių taikymas tyrimuose

„Voyant Tools“ platforma, integruojanti kiekybinius teksto analizės metodus, sukurta skaitmeninės humanitarikos komandos, kuriai vadovavo Stéfanas Sinclairas ir Geoffrey Rockwellas. Tai yra atvirosios prieigos (GitHub, n. d.), lengvai naudojamas ir įsisavinamas įrankis, skirtas didelės apimties tekstų skaitymui ir analizei ir leidžiantis dirbti su įvairių formatų (TXT, HTML, XML, PDF, RTF, MS Word) tekstais bei tekstų kolekcijomis. „Voyant Tools“ tyrėjams padeda analizuoti tekstynus: nustato dažniausiai pasikartojančius žodžius, asociatyvinius žodžių ryšius (kolokacijas), išskiria pagrindines temas, vizualizuoja gautus rezultatus (Sinclair & Rockwell, 2016).

„Voyant Tools“ taikomi įvairių technikų ir metodų deriniai, pvz., „Cirrus“ aplinkoje pasitelkiama kiekybinė žodžių analizė, nustatanti dažniausiai pasikartojančius tekstyno žodžius; „Links“ aplinkoje vaizduojamas kolokacijų grafikas ir tinklai, parodantys asociatyvinius žodžių ryšius bei diskursus; „Trends“ linijinėje diagramoje pateikiamas statistinis žodžių pasiskirstymas tekstyne arba atskirame tekstyno dokumente; „Topic“ aplinkoje temoms nustatyti naudojamas Latentinis Dirichlė paskirstymo (LDA) modelis, taip pat įvairios atvirojo kodo bibliotekos, vizualizavimo bei paieškos įrankiai (Sinclair & Rockwell, 2016).

„Voyant Tools“ platformos galimybės yra aptartos ne vienoje mokslinėje publikacijoje. Dalyje straipsnių išsamiai supažindinama su „Voyant Tools“ įrankių aplinka ir įrankių naudojimo galimybėmis (pvz., Hetenyi et al., 2019), kituose siūlomi „Voyant Tools“ patobulinimai (Milner, 2017 ir kt.).

„Voyant Tools“ platformos taikymo galimybės yra gana plačios. Įrankis taikytas mokslinių tekstų temoms nustatyti (Miller, 2018; Sampsel, 2018); įvairių autorių (pvz., rašytojų Zoros Neale Hurston ir Richardo Wrighto) tekstų skaitymo galimybėms praplėsti, nustatant dažniausiai pasikartojančius žodžius bei stebint, kaip žodžiai kito skirtingais autorių kūrybos laikotarpiais (Rambsy, 2016). Temų kaita tirta ir JAV publikuojamo žurnalo *The American Archivist* leidiniuose (Daines et al., 2018).

„Voyant Tools“ įvairiai pritaikomas tyrimo duomenų turinio analizei. Pvz., atliekant interviu su moterimis, užsiimančiomis verslu Turkijoje, surinkto tekstyno žodžių tarpusavio ryšių analizė parodė, kad dažniausiai pasikartojančio žodžio *economic* (liet. *ekonomika*) stipriausias ryšys yra su žodžiu *independence* (liet. *nepriklausomybė*). O tai leido pagrįsti prielaidą, jog verslo moterys asmeninę nepriklausomybę tiesiogiai sieja su ekonomine nepriklausomybe (Özsungur, 2019).

J. D. Cortésas-Sánchezas (2018), pasitelkęs *Quacquarelli Symonds'* 2016 m. pasaulio universitetų reitingus, atrinko 248 universitetų svetainėse pateikiamą informaciją apie universitetų misiją. Dažniausiai pasikartojančių žodžių dažniai atskleidė, jog universitetai, pristatydami savo misiją, dažniausiai vartoja žodžius, susijusius su žiniomis (angl. *knowledge*), švietimu (angl. *education*) ir suinteresuotomis šalimis, tokiomis kaip studentai ir dalininkai (angl. *stakeholders*), kuriais dažniausiai yra visuomenė. Šie rezultatai sutapo su kitų tyrėjų, analizavusių privačių bei valstybinių universitetų misijas, išvadomis.

Tyrinėti įvairių talpyklų pateikimo formų (angl. *submission agreements*) tekstai, kuriuose duomenis (nepublikuotus straipsnius, publikacijas, konferencijų, mokymo medžiagą ir kt.) pateikiantys asmenys supažindinami su talpyklos taisyklėmis, leidimais, autorinėmis teisėmis ir kita svarbia informacija. Pateikimo formų skaitymas ir aptikti dažniausiai pasikartojantys žodžiai (angl. *non-exclusive, rights, intellectual property*) padėjo išskirti pagrindines temas: „autorinių teisių leidimai“, „teisių pažeidimai – bendrieji“, „teisių pažeidimai – privatumas“ (angl. *copyright permissions, infringement of rights – general, infringement of rights – privacy*). Žodžių dažnio ir dokumento apimties skaičiavimai parodė, kad pateikimo formos yra gana trumpos (vidutinis žodžių skaičius siekia 282 žodžius, kurie vidutiniškai sudaro 9 sakinius), tačiau dėl jose aptinkamų terminų sudėtingumo joms suprasti reikia aukštojo universitetinio išsilavinimo. Autoriai atkreipė dėmesį, kad formų suvienodinimas ir aiškios formuluotės padėtų duomenis pateikiantiems asmenims susipažinti su taisyklėmis ir reikalavimais (Rinehart et al., 2017).

Pasitaiko ir kritinio pobūdžio pastebėjimų. Antai šešių kompanijų pardavimų ir reklamos darbuotojų atsakymus į tyrimo anketų klausimus išnagrinėję G. Hetenyi et al. (2019) pastebėjo, kad labai svarbu atkreipti dėmesį į „Voyant Tools“ įrankio nustatymus, kurie gali iškreipti rezultatą. Nors naudojant šį įrankį gauti rezultatai labai reikšmingi preliminariai analizei ir hipotezei formuluoti, visgi jie turėtų būti atidžiai vertinami ir interpretuojami, kadangi kiekybinės kokybinių duomenų analizės metu prarandami dideli informacijos kiekiai. Gauti rezultatai turėtų būti patikrinti tradiciniais turinio analizės metodais (Hetenyi et al., 2019: 402).

Šio straipsnio problematiką bene geriausiai atliepia G. Spitale (2020) tyrimas. Autorius atkreipia dėmesį, kad mokslinių publikacijų gausa dominančiomis temomis tapo sunkiai aprėpiama ir perskaitoma. Sekti naujausias mokslines publikacijas, ypač tarpdisciplininėmis temomis, tapo iššūkiu. Todėl, siekiant susipažinti su naujausiais tyrimais ir gautais rezultatais, kyla poreikis taikyti kitokias skaitymo ir publikacijų atrankos metodologijas. Siūloma pradėti nuo naujų mokslinių publikacijų paieškos, didelių tekstų kiekio skaitymo strategijos, greitai, efektyviai ir nešališkai atrenkant ir skaitant didelius tekstynus. Autorius pateikia pavyzdį, kaip reikėtų atrinkti bioetikos tema paskelbtų straipsnių publikacijas duomenų bazėse, o surastoms mokslinių straipsnių santraukoms taiko žodžių dažnio analizę. Gauti rezultatai gali parodyti, pvz., paieškos frazės vartojimo publikacijose populiarumą pagal metus (Spitale, 2020: 3–4). Autorius daro išvadą, kad tinkamai pasirinkta paieškos strategija, žodžių dažnio analizė ir reikšminių teksto struktūrų sukūrimas gali padėti susidoroti su didžiuliu informacijos srautu. Skirtingai nei G. Spitale (2020) atliktame tyrime, šiame straipsnyje aptariamos platesnės informacijos apdorojimo galimybės, taikant ne tik dažniausiai pasikartojančių žodžių, bet ir diskursų (dažniausiai pasikartojančių sąvokų ryšių) atranką kokybinei analizei, išsamiau pristatomos interaktyvaus skaitymo galimybės „Voyant Tools“ aplinkoje.

Tekstyno sudarymas ir vizualizavimas

Parengiamasis mūsų tyrimo etapas buvo analizei skirtos mokslo straipsnių kolekcijos sudarymas. Skaitmeninės humanitarikos tematikai skirtus straipsnius atrinkome pasitelkę duomenų bazėse integruotus paieškos įrankius. Atrankai naudojome raktinių žodžių junginius, kurie, mūsų požiūriu, galėtų bendrame tekstų sraute išskirti su skaitmeninės humanitarikos problematika susijusius atskirų humanitarinių mokslų disciplinų straipsnius: „digital anthropology“, „digital ethnography“, „digital history“, „digital archaeology“, „digital literature“, „digital linguistics“ ir „computer linguistics“[9], „digital philosophy“, „digital theology“[10]. Tokiu būdu buvo atrinkti 404 mokslo straipsniai anglų ir kitomis užsienio kalbomis[11].

Norint tekstų analizei taikyti „Voyant Tools“ įrankius, reikia pasirinktus tekstus įkelti į platformą. „Voyant Tools“ platformoje numatyti trys tekstų įkėlimo būdai: internetinių straipsnių nuorodų įkėlimas, teksto įkopijavimas rankiniu būdu arba jo įkėlimas (angl. *upload*) iš kompiuterio disko, kurį ir pasirinkome tekstynui sudaryti[12]. Įkelti tekstai apdorojami ir sukuriamas tekstynas, tinkamas kiekybinei teksto analizei: tekstų žodžiai suvedami į bendrines formas ir pašalinami funkciniai žodžiai, kurie nėra reikšmingi teksto analizei ir vizualiai gali trukdyti pamatyti dažniausiai pasikartojančius žodžius. Pirmoji procedūra atliekama automatiškai, o antrąją galima pakoreguoti, sukuriant papildomą atmetinių sąrašą (angl. *stop words*)[13]. („Voyant Tools“ apdoroja anglų, lietuvių ir daugelį kitų kalbų.) 1 pav. matome „Voyant Tools“ aplinką, kurioje atsiveria automatiškai (mašininiu būdu) išanalizuotas ir vizualizuotas tekstynas.

Atsivėrusiuose languose pateikiami mašininės teksto analizės rezultatai, išreikšti skirtingais būdais. Pažvelgę į 1 paveikslą, rasime penkis langus – *Cirrus*, *Reader*, *Trends*, *Summary* ir *Contexts*. Dauguma jų suaktyvina konkrečias kiekybinės teksto analizės ir vizualizavimo programas, integruotas į „Voyant Tools“ platformą.

Apačioje kairėje esančiame „Summary“ lange pateikiama apibendrinta tekstyno statistika. Matome, kad mūsų skaitmeninės humanitarikos problematikos tekstyną sudaro 404

[9] Pasirinkti tik du kompiuterinės lingvistikos tyrimo lauką atspindinčių raktinių žodžių junginiai, tačiau jų variacijų yra daug dėl plataus tyrimų problematikos lauko ir įdirbio šioje srityje.

[10] Kadangi pasirinktas analizuoti humanitarinių mokslų laukas, paieškos raktiniai žodžiai sudaryti pasitelkiant mokslų krypčių klasifikatorių (2019), kuriame prie humanitarinių mokslų srities (H 000) priskirtos kryptys: filosofija, teologija, menotyra, filologija, istorija ir archeologija, etnologija. Kadangi etnologijos kryptis pasaulinėje literatūroje dažnai keičiama antropologija, papildomai įtraukėme ir antropologijos terminą. Menotyros krypties atsisakyta dėl skirtingos ir labai plačios tyrimų problematikos, jai būtų reikalinga atskira analizė.

[11] Iš 404 straipsnių 350 yra anglų ir 54 straipsniai kitomis užsienio kalbomis (36 ispanų, 5 prancūzų, 3 rusų, po 2 portugalų, lenkų, italų, po 1 serbų, ukrainiečių, korėjiečių, slovakų). Atliekant straipsnių atranką pagal pasirinktus raktinius žodžius, straipsniai kitomis nei anglų kalbomis atrinkti dėl angliškų terminų, vartojamų santraukose anglų kalba.

[12] Galima naudotis internetine „Voyant Tools“ versija arba atsisiųsti į kompiuterį „VoyantServer“. Mūsų atveju atrinktų straipsnių analizė atlikta lokaliame serveryje, kadangi „Voyant Tools“ riboja įkeliamų dokumentų kiekį (žr. prieigą per internetą: https://voyant-tools.org/docs/#!/guide/server).

[13] Sukūrėme papildomą 493 pozicijų atmetinių sąrašą, į kurį įtraukėme tekstyne pasirodančius nereikšmingus žodžius, skaičių fragmentus, skaičių ir žodžių junginius (pvz., 00, 00008, 01, 023x, id, ieee, iitg, liub, s2212, s8755, sbref0025, sci, swj490, tr, cu, 193, 156 ir kt.).

1 pav. **Tekstyno vizualizacija „Voyant Tools“ aplinkoje**

straipsniai, 3 688 049 žodžiai ir 240 160 unikalių žodžių formų. Dažniausiai pasikartoja žodis „skaitmeninis“ (angl. *digital*) – 10 392, „duomenys“ (angl. *data*) – 8 898 ir „tyrimas“ (angl. *research*) – 7 230 kartų. Lange „Cirrus“ (kairėje viršuje) matome tuos pačius dažniausiai pasikartojančius žodžius, išreikštus daugumai pažįstama žodžių debesies forma. Lange „Reader“ (viršuje viduryje) atsiveria visi analizuojamų straipsnių tekstai; viršutiniame dešiniajame lange esantis „Trends“ grafiškai išreiškia dažniausiai pasikartojančius raktinius žodžius arba jų junginius pasirinkto žodžių intervalo kontekste. Grafiko horizontalioje ašyje surašyti tekstyno straipsnių pavadinimai, o vertikalioje – kiekviename straipsnyje pasikartojančių raktinių žodžių arba jų junginių skaičius. Apačioje dešinėje matomas langas, kuriame rodomas pasirinktų raktažodžių kontekstas, kurį galima plėsti išskleidžiant žodžių eilutes į abi puses nuo raktažodžio. „Voyant Tools“ platformos programėlės yra susijusios tarpusavyje. Pvz., jeigu lange „Trends“ kompiuterio pelės kairiuoju klavišu paspaustume pasirinkto „bokštelio“, rodančio tam tikro raktinių žodžių junginio pasikartojimo dažnį konkrečiame straipsnyje, viršūnę, tai lange „Contexts“ atsivertų visos eilutės, kuriose šis žodžių junginys pasirodo, o lange „Reader“ spalvotas žymeklis parodytų šių eilučių vietą tekste. Kiekvieno aptarto lango viršutinėje juostoje galime rasti daugiau parinkčių, kurias pasitelkus atsiveria daugiau teksto analizės galimybių.

Šiame straipsnyje pristatysime tik dalį „Voyant Tools“ platformos funkcijų. Pirmiausia – tas, kurių reikėjo tyrimo uždaviniams spręsti. Norint susipažinti su kitomis „Voyant Tools“ platformos funkcijomis, į pagalbą siūlytume pasitelkti jo kūrėjų parengtą metodinę priemonę (Sinclair & Rockwell, 2016).

Tekstyno problematikos apžvalga pasitelkiant statistinę analizę

„Voyant Tools“ atlieka statistinę teksto analizę, suskaičiuodamas dažniausiai pasikartojančias sąvokas visame tekstyne ir parodydamas jas eilės tvarka pagal dažnumą (žr. 2 pav.). Dažniausiai pasikartojantys žodžiai atspindi dominuojančią tekstyno tematiką. Pvz., pirmieji trys dažniausiai aptinkami mūsų atrinkto tekstyno žodžiai parodo, kad jis susijęs su skaitmenine problematika (angl. *digital*), duomenimis (angl. *data*) ir tyrimais (angl. *research*). Tarp kitų žodžių, kurie tekstyne pasikartoja nuo 7 018 iki 1 785 kartų, randame platesnę tekstyno problematiką atskleidžiančias sąvokas. Platesnį atrinkto tekstyno turinį atspindi žodžiai „medijos“ bei „virtuali / internetinė erdvė“, „analizė“, „socialinis“, „kultūrinis“, „procesas“, „bendruomenė“, taip pat – „universitetas“, „studentai“, „mokslas“ ir kt. Dažniausiai pasikartojančių žodžių sąraše aptinkami ir mokslo tyrimų kryptis pristatantys žodžiai – „istorija“, „archeologija“, „literatūra“, „kalba“ – bei įrankį nusakantis žodis „3D“.

Most **frequent words** in the corpus: digital (10392); data (8898); research (7230); social (7018); new (5481); information (5074); use (4938); media (4720); based (3754); online (3728); analysis (3648); used (3625); study (3560); time (3380); using (3249); work (3058); different (2952); students (2854); history (2807); university (2807); studies (2667); cultural (2652); archaeological (2641); model (2621); learning (2598); technology (2436); process (2401); knowledge (2375); 3d (2331); science (2292); public (2231); users (2219); example (2211); case (2208); language (2142); people (2093); literature (2049); context (2041); approach (2003); archaeology (2002); results (1966); internet (1958); text (1958); project (1955); computer (1882); user (1878); community (1877); international (1869); world (1853); available (1847); development (1833); design (1831); education (1825); space (1808); methods (1805); way (1802); like (1791); number (1787); press (1785)

2 pav. **Dažniausiai pasikartojantys tekstyno žodžiai**

3 pav. **Dažniausiai pasikartojančių sąvokų vizualizavimas.**

Paveiksle dažniausiai pasikartojančios sąvokos (vizualiai didesnės ir parašytos vienodos spalvos fone) tekste aptinkamos iki 5-ių žodžių atstumu viena nuo kitos

Tekstyno apžvalgą tęsiame pasitelkę papildomą programėlę „Links“ (prieinamą aktyvuojant prie „Cirrus“ esančias naujas parinktis). Atsidarius „Links“, programa automatiškai atrinko tris dažniausias tekstyno sąvokas (*digital*, *data* ir *research*) ir susiejo jas su kitomis sąvokomis (žr. 3 pav.). Kaip numatyta, ryšiai tarp sąvokų parodomi tuo atveju, kai sąvokos tekste yra viena greta kitos arba ne didesniu kaip 5 žodžių atstumu viena nuo kitos. (Šiuos programos nustatymus galima keisti eksperimentuojant su kiekvienu konkrečiu tekstynu ir ieškant tinkamiausio rezultato.) 3 pav. matome dažniausiai tekstyne pasikartojančių žodžių (*research*, *data*, *digital*) dažniausius ryšius su kitais žodžiais. Šie ryšiai nusako tekstyno diskursus.

Vertinant algoritmine prasme, „Links“ įrankis sudarė žodžių krepšį (aptartą anksčiau) ir vizualizavo, naudojant atitinkamas vizualizacijos priemones. Tinkamas vizualizavimo priemonių parinkimas leidžia gerokai palengvinti analizės rezultatų suvokimą. Tekstynų analizėje labai dažnai naudojamas „WordCloud“ tipo vizualizavimas (dažniau aptinkami žodžiai vaizduojami didesnio formato, rečiau aptinkami – mažesnio), kurio vienas iš variantų naudojamas toliau pateikiamose iliustracijose.

Interpretuojant 3 paveikslą, nesunku pagrįsti mintį, kad mūsų tyrinėjamo tekstyno pagrindinė problematika yra susijusi su skaitmeniniais tyrimais ir duomenimis, dažnos temos yra duomenų rinkimas (angl. *data–collection*), duomenų analizė, duomenų panaudojimas. Kitos susijusios temos būtų socialiniai tyrimai ir tyrimų metodai, o su skaitmenine (angl. *digital*) tematika susieti diskursai yra skaitmeninė humanitarika, skaitmeninė archeologija, skaitmeninė istorija ir skaitmeninis raštingumas.

4 pav. **Ryšių tarp sąvokų stiprumo vizualizavimas („Links“).**

Sąvokas jungiančių linijų storis rodo ryšių tarp sąvokų pasikartojimo dažnumą tekste. Kuo linija storesnė, tuo dažniau ja sujungti žodžiai tekste aptinkami penkių žodžių atstumu vienas nuo kito. Užvedus pelės žymeklį ant norimos sąvokos, programa išryškina jos dažniausiai pasikartojančius ryšius su kitomis sąvokomis. Pavyzdyje matome paryškintus *digital* ryšius su *humanities*, *literature*, *theology*, *ethnography*, *archaeology*, *anthropology* ir *history.*

Programa leidžia peržiūrėti ir sąvokų ryšių pasikartojimo tekste dažnumą. Pvz., spragtelėję pele ant sąvokos *digital*, pamatome daugiau sąvokų ir jų ryšių (žr. 4 pav.). 4 paveiksle pavaizduoti žodžio *digital* ryšiai atspindi paieškos duomenų bazėse vartotus raktažodžių junginius, tačiau išryškėja ir kitos temos, t. y. *digital* ryšys su *humanities*, *media*, *data*, *research*, *literacy*.

Giliau patyrinėję tekstyną „Links“ įrankio pasiūlytu „medžio“ principu, galime bandyti atsekti temų „grandines“ (žr. 5 pav.).

5 pav. **Diskursų paieškos vizualizacija („Links“)**

5 paveiksle pristatytas sąvokų medis parodo dažniausiai pasikartojančius reikšmingų sąvokų ryšius nuo stambiausių iki mažesnių šakelių:

Digital{literacy{skills, teachers, students}}
Digital{libraries{archives (collections<service>), big, data, information}}
Digital{tools{support, techniques, methods, model, analysis, research}}
Digital{collections{archives, libraries, services, research}}
Digital{cultural{industries, heritage (archeology<applications>), studies (social, history, research)}}
Digital{ethnography{qualitative, approach, virtual, internet, online, social, media, spaces, anthropology, research}}
Digital{archeology{heritage (cultural), applications (systems, methods)}}

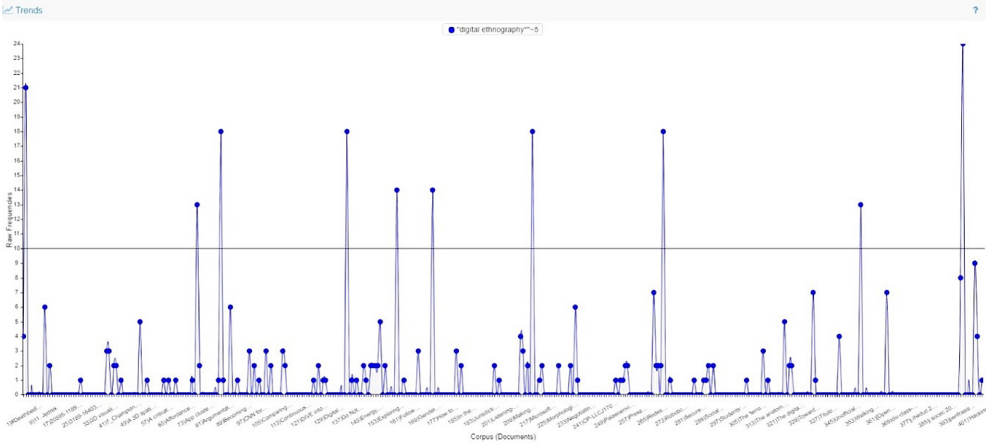
Digital{literature{digitalization, research, culture, review}}
Digital{history{literature, research, studies, education}}
Digital{humanities{sciences, arts, theology, new, social, research, data, scholarship, studies, computing}}

Peržiūrėję gautus rezultatus, galime išskirti tekstyne pastebėtus diskursus: skaitmeninio raštingumo įgūdžių problema; bibliotekų archyvų skaitmeninimo ir skaitmeninių kolekcijų problematika, apimanti ir didelių informacijos kiekių apdorojimą; skaitmeniniai įrankiai ir metodai bei reiškinių modeliavimas; skaitmeninis kultūros paveldas, susijęs su archeologija, visuomenės ir istorijos tyrimais, sisteminių tyrimų metodų taikymas; skaitmeninė etnografija, įskaitant socialinių medijų tyrimus; skaitmeniniai metodai archeologijoje; skaitmeninė literatūra, skaitmeninimas ir tyrimai; skaitmeniniai istorijos tyrimai ir studijos. Straipsniuose taip pat aptikti teoriniai svarstymai apie skaitmeninę humanitariką ir jos ryšį su mokslu, menais, teologija ir socialiniais tyrimais.

Aptarti diskursai atskleidžia skaitmeninės humanitarikos veiklas, būdingas humanitarinių mokslų disciplinoms, ir, savo ruožtu, apibūdina aktualų skaitmeninės humanitarikos interesų ir veiklos lauką.

Diskursų atranka kokybinei analizei

Apžvelgę tekstyno problematiką, galime pasirinkti mus dominančius tekstyno diskursus tolesnei analizei. Jeigu norime išsamiau paskaityti straipsnius, reikia surasti būdą, kaip iš didelio tekstyno atrinkti nedidelį labiausiai tematiką reprezentuojančių straipsnių kiekį. Atrankai pasitelkėme įrankį „Trends“, parodantį dažniausiai tekstyne vartojamų žodžių linijinę diagramą, kuriai sudaryti naudojamas *tf-idf* koeficiento skaičiavimas. Į paieškos langelį įvedę pirmuosius du vieno iš atrinktų diskursų žodžius, pvz., „digital ethnography“, gavome šių žodžių ryšių pasikartojimo dažnių atskiruose straipsniuose grafiką (žr. 6 pav.).



6 pav. **„Digital ethnography“ diskursą pristatančių straipsnių kreivė („Trends“).**

Grafiko horizontaliojoje ašyje surašyti tekstyno straipsnių eilės numeriai ir straipsnių pavadinimai, o vertikaliojoje – skaičiai, rodantys paieškos žodžių junginio „digital ethnography“~5 (nuo

kiekvieno žodžio 5 žodžių atstumas į abi puses) pasikartojimų dažnį kiekviename tekstyno dokumente. Aukščiausiai iškylantys kreivės taškai rodo straipsnius, kuriuose paieškos žodžiai dažniausiai pasikartoja. Grafike horizontalus brūkšnys, nubrėžtas straipsnio autorių ties 10 kartų pasikartojančių paieškos žodžių riba, atidalija didžiausio problematikos dažnio straipsnius.

Iš viso buvo surasti 95 straipsniai, kuriuose paieškos žodžių ryšiai pasikartojo nuo 1 iki 24 kartų. Kadangi mūsų tikslas buvo atlikti išsamesnę reprezentatyvių straipsnių turinio analizę, pasirinkome tik tuos straipsnius, kuriuose paieškos žodžių junginys „digital ethnography“~5 pasikartoja dešimt ir daugiau kartų. Atrinkome 10 straipsnių, iš kurių sukūrėme naują patekstynį (angl. *subcorpus*), kurį būtų patogu aprėpti taikant interaktyvųjį skaitymą[14].

Interaktyvusis skaitymas

Interaktyviuoju skaitymu vadiname iš karto kelių „Voyant Tools“ platformos įrankių naudojimą pereinant nuo statistinių patekstynio vizualizacijų prie tekstų fragmentų peržiūrėjimo (įrankis „Contexts“) ir sugrįžtant į originalų tekstą.

Jeigu atrinkto išsamesnei analizei patekstynio „digital ethnography“ tyrimą pradedame nuo žodžių debesies vizualizacijos (įrankis „Cirrus“), galime iškelti prielaidą, kad pagrindinė tekstyno problematika sietina su skaitmeninių medijų tyrinėjimais, o vienas svarbesnių tyrimo klausimų yra visuomenės narių identiteto tyrimai (žr. 7 pav.). Pažiūrėję atidžiau, randame ir tyrimo dalyvius – komandas, grupių narius, gerbėjus, interneto vartotojus. Matome, kad dominuojanti skaitmeninės etnografijos tyrimų erdvė yra „Twitter“ platforma, pokalbių programėlė „WhatsApp“.

7 pav. **Patekstynio „digital ethnography“ žodžių dažnio vizualizacija.**

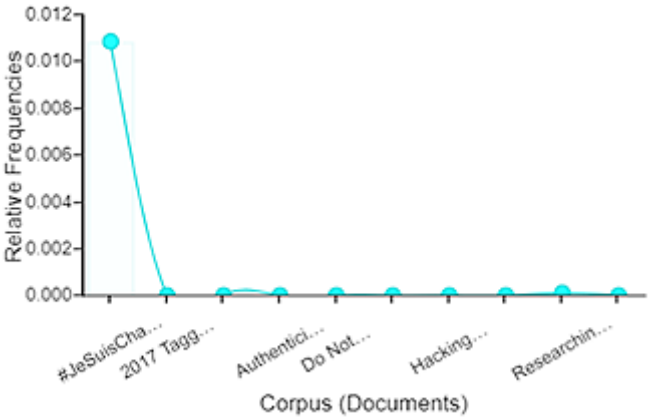
Paveiksle didžiausiu šriftu parašyti žodžiai reiškia dažniausią jų pasikartojimą tekste („Cirrus“).

[14] Sudarytą skaitmeninės etnografijos patekstynį sudaro 10 straipsnių. Prieiga per internetą: https://voyant-tools.org/?stopList=keywords-053ad22d909c37163a7af0132a6a100e&panels=cirrus%2Creader%2Ctrends%2Csummary%2Ccontexts&corpus=8f64e2173a9d968178011dc6b2c1ee2e [žiūrėta 2023 m. sausio 11 d.].

Norėdami išsamiau patyrinėti surastą skaitmeninių medijų problematiką, pereiname į „Links“ aplinką. Įvedę sąvoką „media“ į „Links“ įrankio langą, pamatysime naujus sąvokų ryšius – medijų tyrimai, skaitmeninės medijos, medijų komunikacija ir medijų įvykiai. Savo ruožtu, suaktyvinę, pvz., sąvoką „communication“, galime išskleisti su ja susijusias sąvokas „issue“, „new“ ir „media“ (žr. 8 pav.). Algoritmine prasme matome vizualizuotą žodžių krepšį.

8 pav. **Sąvokos „media“ ryšiai („Links“)**

Norėdami išsamiau peržiūrėti dažniausių sąvokų ryšių kontekstą konkrečiuose straipsniuose, lange „Trends“ įvedame „media* events*“~5[15] ir atsirenkame norimus straipsnius (žr. 9 pav.). Šiuo atveju kompiuteris ieško *tf-idf* koeficiento ir gautas vertes panaudoja grafikui sudaryti.



9 pav. **Sąvokų ryšio „media events“ dažnis atskiruose straipsniuose.**

Paveikslo horizontaliojoje ašyje surašyti straipsnių pavadinimai, o vertikaliojoje – sąvokos „media events“ pasikartojimo dažnis kiekviename straipsnyje („Trends“).

[15] Kitas būdas pamatyti pasirinktas sąvokas atskiruose straipsniuose – pele paspausti žodžių ryšį vaizduojančią liniją „Links“ įrankio aplinkoje, tuomet lange „Trends“ parodomas žodžių „media events“ pasikartojimas atskiruose dokumentuose. Šiuo atveju pasirinktų sąvokų nereikia įvesti ranka.

Matome (9 pav.), kad tema „media events“ (liet. *medijos įvykiai*) buvo surasta tik viename iš dešimties patekstynio straipsnių. Atvėrę dominantį dokumentą „Voyant tools“ lange „Contexts“[16], rasime, kad minėtas „media events“ diskursas išryškėja J. Sumialos et al. (2016) publikacijoje, kurioje analizuojami hibridiniai medijos įvykiai (angl. *hybrid media events*). Pasirinkę išsamiai perskaityti straipsnį, matome, kad autoriai, pasitelkę Charlie Hebdo[17] įvykių analizę *Twitter* platformoje, sukūrė hibridinių medijos įvykių tyrimo metodologiją, kurioje derinami skaitmeninės etnografijos, automatinio turinio analizės, socialinių tinklų analizės ir kokybinio tyrimo metodai (Sumiala et al., 2016: 99). Autorių pasiūlyta metodologija padeda geriau suprasti hibridinės medijos įvykius, jų simbolinę išraišką ir reikšmes.

Plačiau nagrinėdami dažniausių sąvokų ryšių kontekstą patekstynio „digital ethnography“ straipsniuose, lange „Trends“ galime toliau rinktis mus dominančius diskursus ir juos atidžiau tyrinėti. Pvz., pažvelgę į dažniau pasitaikančias sąvokas, pasirenkame išsamiau panagrinėti „stipriausią“ ryšį turinčias sąvokas „team“ (liet. *komanda*), kuri patekstynyje pasikartoja 343 kartus, ir „identity“ (liet. *tapatybė*), kuri pasikartoja 321 kartą (žr. 10 pav.).

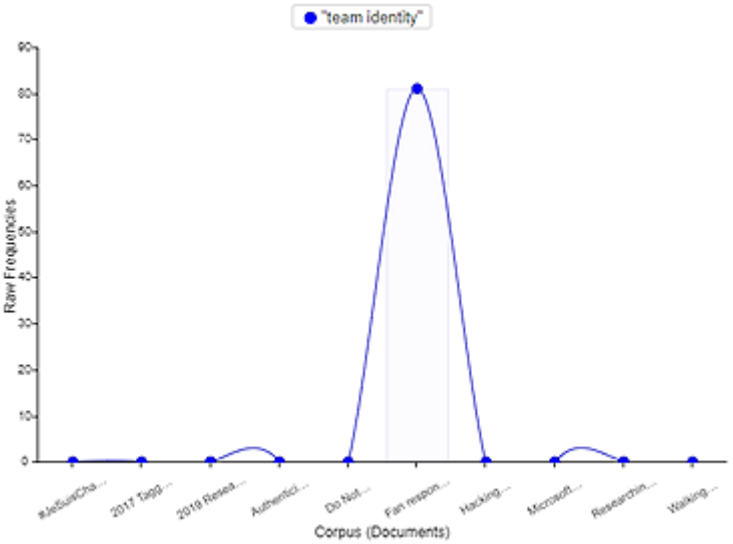
10 pav. **Sąvokos „team*“ ir „identity“ ryšiai („Links“)**

Diskursas „team identity“ (liet. *komandos tapatybė*) daugiausiai kartų pasikartoja C. E. Wegner et al. (2019) straipsnyje, kuriame analizuojama futbolo komandos gerbėjų tapatybė (žr. 11 pav.). Pažvelgę giliau randame, kad autoriai pasitelkia skaitmeninės etnografijos metodą diskusijoms internete stebėti. Autoriai pastebi, kad Nacionalinės futbolo lygos (NFL) „Rams“ gerbėjai, reaguodami į futbolo komandos veiklos vietos pasikeitimus (kai komanda persikėlė iš Sent Luiso, Misūrio valstijoje, į Los Andželą,

[16] Paspaudę pele grafiko (9 pav.) bokštelio viršūnę, programos lange „Contexts“ galėsime peržiūrėti teksto eilutes su mums rūpimais teksto fragmentais. Visą tekstą galime skaityti „Voyant Tools“ programos lange „Reader“.

[17] Šiuo pavadinimu įvardijami 2015 m. Paryžiuje įvykę teroristiniai išpuoliai prieš žurnalo „Charlie Hebdo“ redakciją. Vienas iš žymiausių šio įvykio išraiškų – šūkis „Je Suis Charlie“ (pranc. *Aš esu Charlie*), simbolizuojantis solidarumą ir saviraiškos laisvę (Sumiala et al., 2016: 98).

Kalifornijos valstijoje), keitė ir savo tapatybės reikšmę bei prasmę, kartais išreikšdami pasipriešinimą vykstantiems pokyčiams.



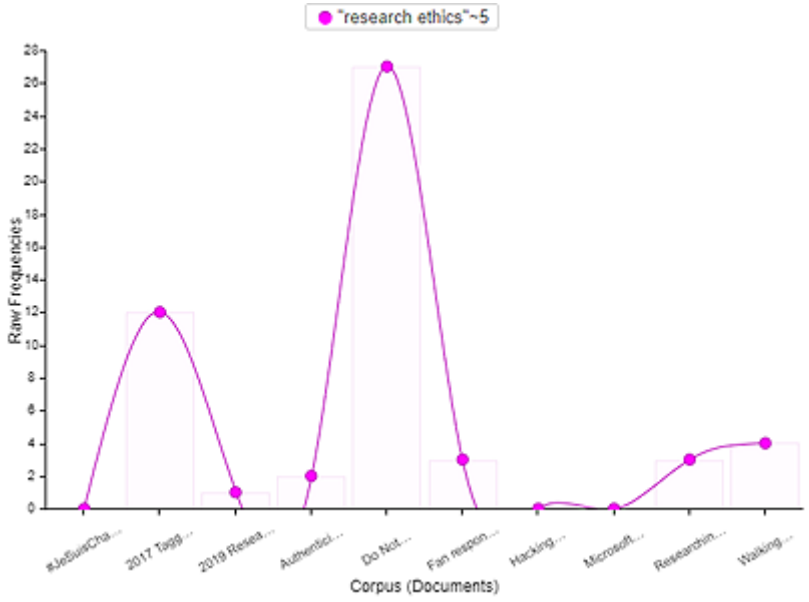
11 pav. **Sąvokų „team identity“ dažnis atskiruose straipsniuose („Trends“)**

Ieškodami dominančių diskursų, išsamesnei analizei galime rinktis ne tik stipriausius ryšius turinčias sąvokas, bet ir panagrinėti kitus, pvz., „skaitmeninių pėdsakų“ (angl. *digital traces*) ar „skaitmeninės erdvės“ (angl. *digital space*), diskursus (žr. 12 pav.).

12 pav. **Sąvokos „digital*“ ir „traces“, „digital*“ ir „space“ ryšiai („Links“)**

„Skaitmeninių pėdsakų“ diskursas matomas K. Møllerio ir B. Robardso (2019) publikacijoje, kurioje autoriai aptaria „mažų duomenų“ socialinių tinklų kokybinio tyrimo metodus[18]. Jų tarpusavio derinimas padeda pamatyti kasdienius naudojimosi socialiniais tinklais aspektus, kurie nublanksta tiriant platų didelės apimties medijų mobilumo kontekstą.

„Skaitmeninės erdvės“ diskursas dažniausiai pasikartoja N. Crowe ir K. Hoskins (2019) straipsnyje, kuriame analizuojami jaunimo andergraundiniai internetiniai judėjimai „Ana Girls“ ir „Pro-Ana“ (*Pro-Anorexia*). Taikant ilgalaikį skaitmeninės etnografijos metodą, stebėti judėjimų internetinio forumo įrašai, blogai, tinklalapiai. Galime prieiti prie išvados, kad diskusijos skaitmeninėje erdvėje prisideda įvertinant žmonių, turinčių valgymo sutrikimų, reiškinio mastą ir jų teisėto gyvenimo būdo pasirinkimą.



13 pav. **Sąvokų „research ethics“ dažnis atskiruose straipsniuose („Trends“)**

Anksčiau minėti diskursai aptikti pavienėse publikacijose, o „tyrimo etikos“ (angl. *research ethics*) problematika pasitaiko dažniau (žr. 13 pav.). Ji siejama su klausimu, kaip apsaugoti asmenų privatumą naudojant privačiose diskusijose gautus duomenis, pvz., var-

[18] Išskirti metodai: 1. „Žingsnis po žingsnio“ metodas (angl. *walkthrough method*), taikomas norint susipažinti su programa, jos veikimo principu bei galimybėmis. 2. „Išplėstinio medijų“ metodo (angl. *media go-along method*) metu, derinant interviu ir dalyvaujamąjį stebėjimą, siekiama suprasti, kaip tiriamieji naudojasi asmeninėmis mobiliomis priemonėmis, socialiniais tinklais. 3. „Slinkties atgal“ metodu (angl. *scroll back method*) tyrėjas kartu su tyrimo dalyviais, stebėdami asmeninius profilius ar socialinių tinklų laiko juostas, fiksuoja pokyčius, aiškinasi įvykius, komentuoja pasikeitimus. Pirmasis ir antrasis metodai orientuoti į vartotojų naudojimosi programomis patirtis, trečiasis – į skaitmeninius pėdsakus, paliktus laiko juostos istorijoje. Pirmasis metodas fiksuoja tyrėjo patirtį, kuomet susipažįstama su programos veikimo principu, antrasis ir trečiasis – orientuojasi į tiriamuosius, jų socialinių tinklų naudojimosi patirtį. Ši patirtis yra asmeninė, intymi ir reikalauja tam tikrų etinių sprendimų (Møller & Robards, 2019).

totojų privatumą (Barbosa & Milan, 2019), politinio aktyvizmo narių „WhatsApp“ grupės privačių diskusijų privatumą. Siūloma grįžti prie etikos pagrindų ir laikytis „nekenkimo“ pozicijos, skatinti dialogą, įtraukiant tyrimo subjektus, nuolat informuojant dalyvius apie atliekamą tyrimą ir jo pobūdį ir kt. Tyrėjas, esant galimybei, turėtų pasirinkti savarankiškai veikiančią infrastruktūrą, galinčią apsaugoti vartotojo duomenis. Svarbu užtikrinti visišką tiriamųjų, ypač pažeidžiamų grupės narių, anonimizavimą (vien tik pseudonimų suteikimas ne visada gali apsaugoti tiriamojo tapatybę), derėtų vengti tiesioginio citavimo bei perfrazavimo (Barbosa & Milan, 2019: 58).

C. Tagg et al. (2017) „tyrimo etikos“ klausimus sieja su tyrėjo ir tiriamojo santykiais, siūlo sudėtingesnių bendravimo situacijų sprendimus, kai, pvz., tiriamieji pasidalina asmeniniais duomenimis ir tyrėjas, stengdamasis nepažeisti tiriamųjų privatumo, turi nutarti, kuriuos duomenis naudoti, o kurių atsisakyti. Straipsnio autoriams kyla klausimas, ar tyrėjai yra pakankamai kompetentingi interpretuoti internetines žinutes. Tyrėjas, bendraudamas skaitmeninėje erdvėje, priešingai nei gyvo bendravimo metu, nemato pašnekovo gestų, intonacijų, o žinutėse pateikti jaustukai neatspindi visų norimų išreikšti emocijų. Tyrėjų požiūriu, šiuo atveju internetinę tiriamųjų veiklą gali padėti interpretuoti duomenys, gauti interviu, stebėjimo metu, o kilus neaiškumų verta grįžti atgal pas pašnekovą (Tagg et al., 2017: 10).

Tęsdami tekstų analizę, galime paieškoti kitų mus dominančių diskursų, identifikuoti, fragmentiškai peržvelgti ir / arba nuosekliai perskaityti straipsnius, kuriuose jie atsiskleidžia. Šis struktūruotas skaitymo būdas padeda sisteminti problematiką. Svarbu tai, kad, taikydami interaktyviojo skaitymo principą, turime galimybę sugrįžti į išsamų autoriaus tekstą, todėl didėja tikimybė neprarasti kitų svarbių duomenų.

Apibendrindami šį skyrių priminsime, kad tekstyną (ir patekstynį) sudarėme iš duomenų bazių atrinkdami temas pagal raktažodžius; po to – pasitelkę kiekybinį metodą (naudodami „Voyant Tools“ aplinką) nustatėme svarbiausius patekstynio diskursus, kurių reprezentatyvius pavyzdžius atsirinkome išsamiam skaitymui.

Išsamiojo skaitymo rezultatai parodė, kad mūsų pasirinkti diskursai, pasitaikę vienoje ar keliose patekstynio publikacijose, gana tiksliai atspindėjo atskirame straipsnyje nagrinėjamą tyrimo problematiką. Išsamiau panagrinėję keletą diskursų, pamatėme, kad straipsnių autorius domino tyrimo skaitmeninėje erdvėje etikos klausimai, taip pat atliekami skaitmeninės erdvės tyrimai, apimantys tiek pačią skaitmeninę erdvę, medijas, tiek šioje erdvėje veikiančias grupes, jų tapatybę.

Diskusijos

Norime pabrėžti, kad straipsnyje pristatytas teksto analizės pavyzdys nėra universalus receptas, nurodantis, kaip reikėtų dirbti su didelės apimties duomenimis ar skaityti konkrečius tekstus. Kiekvienu konkrečiu atveju analizei galima pasitelkti ne tik pristatytus šiame straipsnyje, bet ir kitus „Voyant Tools“ platformoje įdiegtus įrankius, arba ir kitas tekstinių duomenų analizės programas (anksčiau minėtos „AntConc“, „WordSmith Tools“, „Sketch Engine“). O programavimą įvaldę studentai, dėstytojai ir / arba tyrėjai galbūt

mieliau rinksis *Python*, *R*, *Go* ar kitas programavimo kalbas, kurios ne tik atliktų panašias funkcijas kaip „Voyant Tools“, bet ir leistų tyrėjui pačiam susikurti jam patogiausią tyrimo scenarijų.

Tekstų tyrėjams svarbūs ne vien kompiuterinės teksto analizės įrankiai, o kiekybinių ir kokybinių metodų žinios, kurių pagrindu kuriamos programos. Tyrimo metodų pasirinkimas visuomet priklauso nuo tyrėjų tikslų, pvz., kodėl jiems reikalinga didelės apimties mokslinės literatūros analizė, ką jie joje norėtų surasti ir išsiaiškinti. Pvz., šio straipsnio tikslas buvo, taikant atsietąjį skaitymą, nustatyti skaitmeninės humanitarikos problematiką apimančių straipsnių diskursus, pasirodančius atvirosios prieigos duomenų bazių humanitarinės krypties žurnaluose. Tokios žinios naudingos norintiems orientuotis problematikoje, jos taip pat padeda norint atlikti išsamesnes apžvalgas.

„Voyant Tools“ įrankis sukurtas kiekybinės teksto analizės principu – kompiuterinė programa skaičiuoja pasikartojančių žodžių ir jų junginių dažnius pagal pasirenkamus nustatymus, kuriuos turi prisitaikyti patys tyrėjai. Mums buvo priimtinas programos kūrėjų pasiūlytas 5 žodžių kontekstas. Pasitelkę „Voyant Tools“, turėjome galimybę tekstų analizę atlikti greičiau ir, mūsų požiūriu, patikimiau, nei būtume tai darę tradiciniu būdu, nes kompiuterinė programa, skaičiuodama tekstų žodžius ar nurodydama jų tarpusavio ryšius (ieškodama galimų asociatyvių reikšmių), iš karto sistemina duomenis ir vizualizuoja rezultatus. Savo ruožtu, kiekybinės teksto analizės metodai gali būti naudingi siekiant pačių įvairiausių tikslų, pvz., ieškant tapatybių modelių, galinčių atsiskleisti išsamiųjų interviu, klausimynų, interneto svetainių ir kituose tekstuose, siekiant surasti tyrimui reikšmingus klausimus, formuluojant hipotezes (Miller, 2018).

Paminėjus kompiuterinės analizės privalumus, svarbu atkreipti dėmesį ir į jos trūkumus. Iš vienos pusės, mūsų taikyti mašininiai atrankos metodai padėjo sumažinti analizei reikalingų tekstų kiekį iki konteksto analizei aprėpiamos apimties, tačiau ta pati atranka parodė ir galimas paklaidas. Pvz., įrankis „Trends“ parodo tekstyno dokumentus, kuriuose randa greta esančias sąvokas, pvz., „digital ethnography“, bei tekstų fragmentus, kuriuose „digital“ ir „ethnography“ nutolę vienas nuo kito pasirinktu žodžių atstumu, pvz., „...interviews in a ‚*digital* sensory *ethnography*‘ to explore the...“, „...mobilities in contemporary *digital* ‚small data‘ *ethnography* at the same time...“. Tačiau susidūrėme su atvejais, kai statistika suklaidino, nes tyrinėjamo sąvokų junginio „digital literature“ žodžiai atsidūrė dviejuose skirtinguose sakiniuose, kurių kiekvienas turėjo skirtingą kontekstą, pvz., „... to create a *digital* identity. In scientific *literature*...“. Tokiems netikslumams surasti gali pasitarnauti teksto fragmentų, susijusių su paieškos žodžiais, peržiūrėjimas „Contexts“ aplinkoje. Taip peržiūrėdami sudarytą patekstynį, aptikome, kad neretai žodžių dažnis būna klaidingas dėl to, kad kompiuteris suskaičiuoja ir tuos sąvokų junginius, kurie yra ne pagrindiniame straipsnio tekste, o literatūros sąraše ar straipsnio pavadinime, kuris PDF formato tekstuose dažnai pateikiamas kiekvieno puslapio apačioje.

Kad išvengtume anksčiau įvardytų netikslumų, sudarant tekstyną reikėtų pagalvoti, kaip papildomai išvalyti tekstus. Iš duomenų bazių PDF formatu parsisiųstus straipsnius būtų naudinga paversti į TXT formatą ir ištrinti neesminę informaciją. Tai gali būti, pvz., autorių biografijos, padėkų tekstai, leidyklos pateikiama techninė informacija, literatūros

sąrašas. Didelį kiekį straipsnių gali būti sudėtinga išvalyti, todėl šiam darbui naudinga pasitelkti, pvz., *Python* programavimo kalbą. Tačiau kiekvienu atveju sprendimą turėtų priimti tyrėjas žiūrėdamas specifinių savo interesų. Pvz., mūsų tyrimo atveju buvo nuspręsta literatūros sąrašus palikti, nes mums buvo naudinga matyti, kokiais šaltiniais remiasi straipsnių autoriai.

Kitas momentas, į kurį svarbu atkreipti dėmesį analizuojant tekstus su „Voyant Tools“ platformos įrankiais, yra tai, kad statistinis tekstyno žodžių dažnis nebūtinai reikš jų dominavimą visuose tekstuose. Kitaip tariant, tyrėjo „atrasta“ problematika gali būti būdinga penkiems straipsniams iš penkiolikos arba vienam iš dešimties tuo atveju, jeigu tuose straipsniuose sąvokos pasikartoja dažniausiai. Pvz., mūsų analizuotame patekstynyje pasikartojančios sąvokos „media“, „research“, „digital“, „social“, „data“, „ethnography“, „new“ ir kt. aptiktos beveik visuose skaitmeninės etnografijos patekstynio straipsniuose. Tačiau dalis sąvokų bei jų junginių buvo aptikti tik pavienėse publikacijose. Pasikartojimas gali būti susijęs su autorių kalbos stiliumi arba aprašomuoju teksto pobūdžiu. Norint išvengti klaidingų apibendrinimų, verta tekstyną nagrinėti pasitelkiant bent kelis „Voyant Tools“ įrankius, o esant reikalui, kompiuterinę teksto analizę suderinti su tradiciniu teksto skaitymu.

Patogi „Voyant Tools“ galimybė – analizuojamą tekstyną patyrinėti keliems tyrėjams, patikrinti interpretacijas, paanalizuoti kitus rūpimus klausimus. Tačiau aptinkame ir apribojimą – tekstyno nuoroda aktyvi tol, kol ji paspaudžiama bent vieną kartą per mėnesį (Sinclair & Rockwell, 2016).

Išvados

Šiame straipsnyje, pasitelkę atsietąjį bei interaktyvų tekstų skaitymo ir analizės metodus (įdiegtus „Voyant Tools“ platformoje), nustatėme skaitmeninės humanitarikos problematiką, atsiskleidžiančią humanitarinių mokslų sričiai priskirtose publikacijose, įtrauktose į „Clarivate Analytics Web of Science“ ir „Scopus ScienseDirect“ duomenų bazes (pateiktose iki 2020 m. sausio 24 d.).

Tekstyno *problematikai*, jo *diskursams* nustatyti labiausiai pasitarnavo „Voyant Tools“ aplinkoje integruotas įrankis „Links“, kuris „medžio“ principu pristato tekste dažniausiai pasikartojančių sąvokų ryšius. Nors dalis rezultatų nebuvo netikėti dėl raktažodžių, naudotų sudarant tekstyną (pvz., akademines subdisciplinas nurodantys raktažodžiai „skaitmeninė literatūra“, „skaitmeninė archeologija“, „skaitmeninė istorija“ ir kt.), atliekant kompiuterinę tekstų analizę iškilo naujų temų, kurių nebuvome numatę, pvz., skaitmeninio raštingumo, skaitmeninių bibliotekų, skaitmeninių kolekcijų ir skaitmeninimo, skaitmeninių įrankių ir metodų, skaitmeninio paveldo, skaitmeninės kultūros diskursai, išsamiau atskleidžiantys aktualias skaitmeninės humanitarikos veiklas.

„Voyant Tools“ aplinkoje integruotas įrankis „Trends“ padėjo surasti greitą būdą, kaip iš didelio straipsnių skaičiaus atrinkti reprezentatyvius dominančios problematikos straipsnius ir sukurti naujus tekstynus gilesnei analizei.

Gilesnei atrinktų straipsnių analizei buvo pasitelktas interaktyvus tekstų tyrinėjimo būdas, pereinant nuo statistinės vizualizacijos („Cirrus“, „Links“, „Trends“) prie tekstų

fragmentų peržiūrėjimo („Contexts“) ir visiško sugrįžimo į originalų straipsnį. Toks interaktyvus skaitymas, derinant statistinę teksto analizę su tradiciniu skaitymu, padėjo sisteminti informaciją neprarandant konteksto.

Svarbu paminėti, kad kompiuterinė teksto analizė, grindžiama statistiniais metodais, dėl tyrėjo interpretacijų įgauna kokybinį pobūdį, be to, inspiruoja naujas, netikėtas įžvalgas.

Apibendrindami „Voyant Tools“ vizualines savybes, galime pasakyti, kad kompiuterio ekrane vienu metu matomos skirtingos teksto statistinės analizės išraiškos palengvino patį analizės procesą. Programa pasiūlė naujas tolesnės analizės galimybes, kurių nebuvome numatę pradėdami tekstų tyrinėjimą. Kiekvienas tyrėjas gali pasirinkti jam geriausiai tinkamus analizės įrankius (integruotus į „Voyant Tools“), priklausomai nuo savo turimų žinių apie teksto analizę, turimų kompiuterinio raštingumo įgūdžių ir keliamų tyrimo uždavinių. Pradedantiems tyrėjams gali būti ypač naudinga išbandyti kompiuterinius įrankius, nes šie bandymai leidžia įgyti pradinių žinių, kurios paskatins giliau domėtis teksto analizės metodais.

Padėkos

Dėkojame žurnalo „Information & Media“ vyriausiajam redaktoriui prof. dr. Vladislavui Fominui ir recenzentams už pagalbą rengiant straipsnį spaudai.

Literatūra

Alsmearat, K., Al-Ayyoub, M., & Al-Shalabi, R. (2014). An extensive study of the bag-of-words approach for gender identification of arabic articles. In *2014 IEEE/ACS 11th International Conference on Computer Systems and Applications (AICCSA)* (pp. 601–608). IEEE. https://doi.org/10.1109/AICCSA.2014.7073254

Barbosa, S., & Milan, S. (2019). Do not harm in private chat apps: Ethical issues for research on and with WhatsApp. *Westminster Papers in Communication and Culture*, *14*(1), 49–65. https://doi.org/10.16997/wpcc.313

Bettany-Saltikov, J. (2012). *How to do a systematic literature review in nursing: A step-by-step guide*. McGraw-Hill Education.

Cortés Sánchez, J. D. (2018). Mission statements of universities worldwide: Text mining and visualization. *Intangible Capital*, *14*(4), 584–603. http://dx.doi.org/10.3926/ic.1258

Crowe, N., & Hoskins, K. (2019). Researching transgression: Ana as a youth subculture in the age of digital ethnography. *Societies*, *9*(3), Article 53. https://doi.org/10.3390/soc9030053

Daines III, J. G., Nimer, C. L., & Lee, J. R. (2018). Exploring the American Archivist: Corpus analysis tools and the professional literature. *Journal of Contemporary Archival Studies*, *5*(1), Article 3. https://elischolar.library.yale.edu/jcas/vol5/iss1/3/

Das, B., & Chakraborty, S. (2018). *An improved text sentiment classification model using TF-IDF and next word negation*. arXiv preprint arXiv:1806.06407.

Daud, N. M., Hassan, H., & Aziz, N. A. (2013). A corpus-based readability formula for estimate of arabic texts reading difficulty. *World Applied Sciences Journal*, *21*(1), 168–173. https://doi.org//10.5829/idosi.wasj.2013.21.sltl.2151

GitHub. (n. d.). *Voyant Tools*. Žiūrėta 2020 m. rugpjūčio 28 d., https://github.com/sgsinclair/Voyant

Harris, Z. (1954). Distributional Structure. *Word*, *10*(2/3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hetenyi, G., Lengyel, A. D., & Szilasi, M. D. (2019). Quantitative analysis of qualitative data: Using voyant tools to investigate the sales-marketing interface. *Journal of Industrial Engineering and Management*, *12*(3), 393–404. http://dx.doi.org/10.3926/jiem.2929

Jänicke, S., Franzini, G., Cheema, M. F., & Scheuermann, G. (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In R. Borgo, F. Ganovelli, & I. Viola (Eds.), *Eurographics Conference on Visualization (EuroVis) (STARs)* (pp. 83–103). The Eurographics Association.

Jurafsky, D., & Martin, J. H. (2021). *Speech and Language Processing* (3rd ed.). Pearson. https://web.stanford.edu/~jurafsky/slp3/

Keturakis, S. (2019). Apie skaitymą iš toli ir iš arti. *LOGOS-A Journal of Religion, Philosophy, Comparative Cultural Studies and Art*, *99*, 103–112. https://doi.org/10.24101/logos.2019.34

Kincaid, J. P., Fishburne Jr, R. P., Rogers, R. L., & Chissom, B. S. (1975). *Derivation of new readability formulas (automated readability index, fog count, and flesch reading ease formula) for Navy enlisted personnel.* Research Branch Report 8–75. Chief of Naval Technical Training: Naval Air Station Memphis.

Luhn, H. P. (1957). A Statistical Approach to Mechanized Encoding and Searching of Literary Information. *IBM Journal of Research and Development*, *1*(4), 309–317. https://doi.org//10.1147/rd.14.0309

Miller, A. (2018). Text mining digital humanities projects: Assessing content analysis capabilities of voyant tools. *Journal of Web Librarianship*, *12*(3), 169–197. https://doi.org/10.1080/19322909.2018.1479673

Milner, M., Wittek, S., & Sinclair, S. (2017). Introducing DREaM (Distant Reading Early Modernity). *DHQ*, *11*(4). http://www.digitalhumanities.org/dhq/vol/11/4/000313/000313.html

Mokslo krypčių klasifikatorius. (2019). *Lietuvos Respublikos švietimo, mokslo ir sporto ministro įsakymu „Dėl švietimo, mokslo ir sporto ministro vasario 6 d. įsakymo Nr. V-93 „Dėl mokslo krypčių ir meno krypčių klasifikatorių patvirtinimo“ pakeitimo“ 2019 m. vasario 20 d. Nr. V-156, Vilnius.* https://e-seimas.lrs.lt/portal/legalAct/lt/TAD/833ca8f2354f11e98893d5af47354b00

Møller, K., & Robards, B. (2019). Walking through, going along and scrolling back: Ephemeral mobilities in digital ethnography. *Nordicom Review*, *40*(s1), 95–109. https://doi.org/10.2478/nor-2019-0016

Moretti, F. (2000). Conjectures on world literature. *New Left Review*, *1*, 54. https://newleftreview.org/issues/ii1/articles/franco-moretti-conjectures-on-world-literature

Özsungur, F. (2019). A research on women's entrepreneurship motivation: Sample of Adana Province. *Women's Studies International Forum*, *74*, 114–126. https://doi.org/10.1016/j.wsif.2019.03.006

Rambsy, K. (2016). Text-Mining Short Fiction by Zora Neale Hurston and Richard Wright using Voyant Tools. *CLA Journal*, *59*(3), 251–258. https://www.jstor.org/stable/44325917

Rinehart, A., & Cunningham, J. (2017). Breaking it down: A brief exploration of institutional repository submission agreements. *The Journal of Academic Librarianship*, *43*(1), 39–48. https://doi.org/10.1016/j.acalib.2016.10.002

Rudžionis, V., Lopata, A., Gudas, S., Butleris, R., Veitaitė, I., Dilijonas, D., Grišius, E., Zwitserloot, M., & Rudžionienė, K. (2022). Identifying Irregular Financial Operations Using Accountant Comments and Natural Language Processing Techniques. *Applied Sciences*, *12*(17), Article 8558. https://doi.org/10.3390/app12178558

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Sampsel, L. J. (2018). Voyant tools. *Music Reference Services Quarterly*, *21*(3), 153–157. https://doi.org/10.1080/10588167.2018.1496754

Sinclair, S., & Rockwell, G. (2016). *Voyant Tools*. https://voyant-tools.org/docs/#!/guide/start

Smith, J. B. (1978). Computer Criticism. *Style*, *12*(4), 326–356. http://www.jstor.org/stable/45108824

Spaulding, S. (1956). A Spanish Readability Formula. *The Modern Language Journal*, *40*(8), 433–441. https://doi.org/10.1111/j.1540-4781.1956.tb02145.x

Spitale, G. (2020). Making sense in the flood. How to cope with the massive flow of digital information in medical ethics. *Heliyon*, *6*(7), Article e04426. https://doi.org/10.1016/j.heliyon.2020.e04426

Stringhini, G., Kruegel, C., & Vigna, G. (2010). Detecting spammers on social networks. In *Proceedings of the 26th annual computer security applications conference* (pp. 1–9). Association for Computing Machinery. https://doi.org/10.1145/1920261.1920263

Sumiala, J., Tikka, M., Huhtamäki, J., & Valaskivi, K. (2016). # JeSuisCharlie: Towards a multi-method study of hybrid media events. *Media and Communication*, *4*(4), 97–108. https://doi.org/10.17645/mac.v4i4.593

Tagg, C., Lyons, A., Hu, R., & Rock, F. (2017). The ethics of digital ethnography in a team project. *Applied Linguistics Review*, *8*(2-3), 271–292. https://doi.org/10.1515/applirev-2016-1040

Van Atteveldt, W., Welbers, K., & Van Der Velden, M. (2019). Studying political decision making with automatic text analysis. *Oxford Research Encyclopedia of Politics*. https://doi.org/10.1093/acrefore/9780190228637.013.957

Wallach, H. (2006). Topic modeling: beyond bag-of-words. *Proceedings of the 23rd international conference on Machine learning*, 977–984.

Wegner, C. E., Delia, E. B., & Baker, B. J. (2020). Fan response to the identity threat of potential team relocation. *Sport Management Review*, *23*(2), 215–228. https://doi.org/10.1016/j.smr.2019.01.001

Zhu, Z., Liang, J., Li, D., Yu, H., & Liu, G. (2019). Hot topic detection based on a refined TF-IDF algorithm. *IEEE access*, *7*, 26996–27007. https://doi.org/10.1109/ACCESS.2019.2893980