10.35539/LTNC.2022.0047.S.R.H.M.I.L.T.70.87

Metadatos balstīta dienasgrāmatu teksta korpusa analīze

Sanita Reinsone, Haralds Matulis,
Ilze Ļaksa-Timinska

Raksts ir tapis Latvijas Zinātnes padomes projektā Nr. lzp-2018/1-0073 veiktā pētījuma rezultātā.

Atslēgvārdi: digitālās humanitārās zinātnes, tāllasījums, teksta analīze, dienasgrāmatu korpuss, autobiogrāfiskie teksti

Ievads

Personiskās dienasgrāmatas ir unikāls laikmeta naratīvais liecinājums, kas laikmetu, notikumus atklāj caur dziļi personisku skatījumu. Nešaubīgi katra dienasgrāmata ir unikāla, jo rakstīta noteiktā vietā un laikā, pierakstos iekapsulējot daudzas autoru šodienas. Tomēr tās visas kopā rada īpašu personisko rakstījumu dimensiju, kur individuālo stilistiku veido autora gaume un rakstīšanas prasme, arī dienasgrāmatu rakstīšanas tradīciju, daiļliteratūras un citu kultūras parādību ietekmes. Atzīstot katras dienasgrāmatas unikalitāti un savpatību, šajā rakstā mēs pievēršamies personiskajām dienasgrāmatām kā kopumam un veicam metadatos balstītu dienasgrāmatu pilotkorpusa analīzi ar mērķi atklāt korpusa īpatnības dažādu kritēriju sastatījumā.

Šī raksta pamatā ir nesena pieredze, kas gūta, veidojot personīgo dienasgrāmatu tekstu korpusu no paša sākuma, proti, sākot ar dienasgrāmatu vākšanu, digitalizāciju un tālāk – materiālu apstrādi korpusa izveidošanai. Šajā rakstā analizējam dienasgrāmatu tekstu korpusa veidošanu, metodoloģiskos pārbaudījumus un sarežģījumus, ar kādiem nācies saskarties, īstenojot šādu iniciatīvu.

Personīgo dienasgrāmatu pētniecībā līdz šim gluži saprotamā kārtā dominējušas kvalitatīvās pieejas. Tās izmantotas arī tad, ja analizējamais materiāls ir bijis ļoti plašs un sarežģīti pārskatāms. Pētnieki apbrīnojamā kārtā spējuši apgūt un sistematizēt pavisam apjomīgas dienasgrāmatas, atrodot un piefiksējot sev interesējošās tēmas un teksta elementus. Kā nesenā publikācijā norāda Džūlija Raka: "[Š] obrīd dzīves pierakstīšanas pētniecības metodes ir, šķiet, pretrunā ar lielo datu metodēm." (Rak 2019: 118) Kaut arī datormetožu izmantošana humanitāro zinātņu pētījumos ir pazīstama jau kopš 20. gadsimta 50. gadiem (Nyhan, Flinn 2016: 1–4; Sula, Hill 2019: 190–206), pēdējās desmitgadēs digitālās humanitārās zinātnes attīstījušās patiesi strauji. No sākotnēji dominējošās datorlingvistikas kā centrālās digitālo humanitāro zinātņu[1] nozares digitālās humanitārās zinātnes ir būtiski paplašinājušās un metodoloģiski sazarojušās, aizvien dziļāk ietiecoties literatūrzinātnes un citu humanitāro zinātņu diskursīvajā laukā.

1 Digitālās humanitārās zinātnes ir salīdzinoši jauns termins, kas tiecas aptvert dažādas humanitāro zinātņu apakšnozares un datormetodes. Iepriekš dominējis jēdziens gan "humanitāro zinātņu programmēšana" (*humanities computing*), gan citi vairāk vai mazāk iekļaujoši vai nozarēm specifiski nosaukumi (Hockey 2004; Unsworth 2004).

Lai gan par datormetožu lietojumu dzīves pierakstīšanas pētniecībā ir paustas zināmas bažas, jo sevišķi saistībā ar autoru privātumu un ētiku (Rak 2019: 117–118), mūsuprāt, autobiogrāfisko materiālu izpētei ar tāllasījuma (*distant reading*) metodēm (Moretti 2013), kas darbojas rokrokā ar kvalitatīvās izpētes paradigmu, ir vērā ņemama perspektīva un labs attīstības potenciāls dzīves pierakstīšanas fenomena izpētē. Jo sevišķi pētnieciski interesantas šādā perspektīvā ir personīgās dienasgrāmatas.

Formāli raugoties, dienasgrāmatas sadalītas daudzās sīkākās vienībās, kur katrai no tām ir atšķirīga metainformācija. Tās ir piesaistītas konkrētam autoram un laikam, kas gluži kā laikrakstu publikācijas, kuras veiksmīgi analizētas un pētītas ar tāllasījuma metodi (Koncar et al. 2020), ļauj laikā pārlūkot un distancēti vērot tēmu un konceptu attīstību, autora noskaņojumu pret noteiktām tēmām un, protams, arī valodas savdabību, variācijas un izmaiņas, ko var vērot diahroniskā attīstībā, ko jo sevišķi ļauj tās dienasgrāmatas, kas nav literāri rediģētas. Tomēr autobiogrāfisko tekstu korpusu veidošana nav vienkārša un ir saistīta ar daudziem sarežģījumiem, ar kādiem nesaskartos, piemēram, periodikas korpusu vai literāro tekstu korpusu veidotāji.

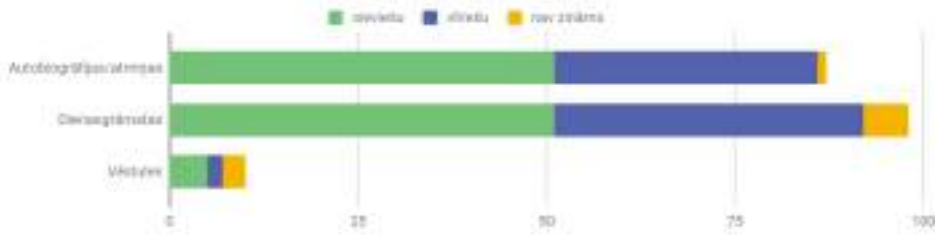
Ceļā uz korpusu – autobiogrāfiju kolekcijas veidošana

Iecere par autobiogrāfisko tekstu korpusu radusies vienlaikus ar Autobiogrāfiju krājuma izveidošanu. Līdz 2018. gadam Latvijā nebija atsevišķas kultūras mantojuma institūcijas, kuras mērķis būtu veidot vienotu un iekļaujošu autobiogrāfisko rakstījumu kolekciju. Tiesa, Latvijas muzeju un bibliotēku krājumos, arī arhīvos un speciālos pētniecisko institūciju krājumos glabājas ievērojama autobiogrāfiskā mantojuma daļa. Tomēr institūcijās glabāto materiālu apjomu un dažādību nav vienkārši precizēt, jo reģistri ir grūti pieejami (tie ne vienmēr ir publiski), tikai daļa no tiem ir digitalizēti un pieejami tiešsaistē[2], un nereti šādi materiāli netiek īpaši piereģistrēti vispār, jo stāv ārpus institūcijas pamatdarbības (piemēram, reģionālajās bibliotēkās).

Turklāt ievērojama autobiogrāfiskā mantojuma daļa nepārtraukti top un glabājas arī privātajos un ģimeņu arhīvos. Šo materiālu autori, autoru pēcteči vai citi īpašnieki tos visbiežāk nenodod glabāšanai kultūras mantojuma institūcijās, jo to glabātājiem ne vienmēr ir skaidrs, kuru institūciju konkrētais materiāls varētu interesēt, kā arī daudzos gadījumos tie oriģinālu vēlas paturēt savā ģimenes arhīvā, nododot mantojumā nākamajām paaudzēm, bet reti kuru mantojuma institūciju interesē uzglabāt digitālu materiālu bez fiziskā oriģināla. Kopumā autobiogrāfiskais mantojums Latvijā vērtējams kā izkaisīts un sadrumstalots. Tomēr tā nav tikai Latvijai vien raksturīga situācija.

Lai līdzsvarotu šo situāciju un autobiogrāfiskajam mantojumam piešķirtu lielāku un redzamāku nozīmi, 2018. gada nogalē LU Literatūras, folkloras un mākslas institūta Latviešu

2 Piemēram, muzeju katalogs un arhīvu datubāze: http://www.nmkk.lv.



1. attēls. LFK Autobiogrāfiju krājuma kolekciju pārskats pēc rakstījuma formas.

folkloras krātuvē[3] tika izveidots Autobiogrāfiju krājums ar mērķi nodrošināt vietu, kur ilgtspējīgi un pieejami saglabāt dažāda veida dzīves rakstījumus neatkarīgi no to valodas, autora reģionālās piederības, reliģijas, politiskajiem uzskatiem, nopelniem politiskajā, sociālajā vai kultūras dzīvē u. tml.

Pirmajos krājuma pastāvēšanas gados lielākā uzmanība tika pievērsta tieši personīgajiem jeb mājas arhīviem, kuros autobiogrāfiskie materiāli ir visvairāk apdraudēti. Lai pievērstu sabiedrības uzmanību Autobiogrāfiju krājumam, veicinātu izpratni par autobiogrāfisko materiālu vērtību un nepieciešamību tos saglabāt, kā arī lai aicinātu sabiedrību iesaistīties un sniegt materiālus digitālai deponēšanai vai nodošanai arhīvam pavisam, tika īstenotas dažādas publicitātes iniciatīvas, tostarp manuskriptu atšifrēšanas digitālās talkas[4], mediju kampaņas un atvērtās arhīva dienas[5], kad ikviens tiek aicināts atnest digitalizēšanai autobiogrāfisku materiālu. Līdz ar to krājuma veidošanas sākumposmā saturu pamatā veido materiāli, ko iesnieguši privātie iesniedzēji. Salīdzinoši nelielu krājuma daļu veido sadarbībā ar mantojuma institūcijām veidotās kolekcijas, kas Autobiogrāfiju krājumam nodotas digitālai deponēšanai.

Pārlūkojot no 2018. gada otrās puses līdz 2021. gada beigām krājumam iesniegtos materiālus (skat. 1. attēlu), redzams, ka aina ir diezgan līdzvērtīga – retrospektīvo formu kolekciju (atmiņu, dzīvesstāstu) krājumā ir tikai nedaudz mazāk nekā dienasgrāmatu kolekciju. Vēstuļu kolekciju niecīgais daudzums izskaidrojams ar to, ka publiskajos uzsaukumos izskanējuši aicinājumi iesniegt dienasgrāmatas un dzīvesstāstus, vēstules nepieminot, lai kolekcijas

3 Latviešu folkloras krātuve ir dibināta 1924. gadā. Tā ir viens no lielākajiem un senākajiem folkloras arhīviem Eiropā, kas šobrīd atrodas Latvijas Universitātes Literatūras, folkloras un mākslas institūta paspārnē. Vairāk informācijas: http://lfk.lv.

4 Regulāri aicinājumi *garamantas.lv* un sociālajos tīklos atšifrēt konkrētus pētniekiem interesējošos materiālus, piemēram, aicinājums šifrēt garās dienasgrāmatas pa fragmentiem (http://garamantas.lv/lv/post/1787/lfk-autobiografiju-krajums-aicina-talka).

5 Piemēram, 2018. gada bija viena no pirmajām "Autobiogrāfiju dienām", aicinot ikvienu atnest materiālus uz Latviešu folkloras krātuvi (https://www.facebook.com/754708391535312/photos/a.754721168200701/761805567492261/).

veidošanas sākumposmā neveidotos pārslodze. Kopskaitā 2021. gada beigās Autobiogrāfiju krājumā iekļautas 207 kolekcijas, kur katru veido viena autora (vai autoru kolektīva) rakstītais. Vairumā no tām dominē viena rakstījumu forma, piemēram, tā ir dienasgrāmata vai dzīvesstāsts. 4,84 % jeb 10 kolekcijās sastopamas vairākas formas.

Tekstuālo kvantitāti šis grafiks gan neparāda. Viena autobiogrāfiskā vienība var būt gan pāris lappušu garš dienasgrāmatas fragments[6], gan desmitgadēm ilgi rakstīta dienasgrāmata[7], kā arī atsevišķos gadījumos vienā kolekcijā var būt iekļauts dienasgrāmatu kopums, kas veidots kā atsevišķs projekts, piemēram, Pandēmijas dienasgrāmatu kolekcija Nr. 166[8], kas ietver 238 autoru dienasgrāmatas. Līdztekus autobiogrāfiskajiem pierakstiem kolekcijās tiek iekļauti arī ar autoriem saistīti papildu materiāli, piemēram, fotogrāfijas. Daļēji publiska piekļuve krājumam tiek nodrošināta LFK digitālajā arhīvā *garamantas.lv*[9].

Liela nozīme krājuma veidošanā ir sabiedrībai, kas tiek iesaistīta ne tikai krājuma veidošanā, bet arī manuskriptu pārveidošanā datorlasāmā formā, izmantojot LFK digitālā arhīva rīkus manuskriptu atšifrēšanai[10]. Digitalizēto manuskriptu pārveidošana datorizētā formā ir viens no svarīgākajiem un arī laikietilpīgākajiem posmiem korpusa un automātiski meklējamas kolekcijas veidošanā, tāpēc brīvprātīgo kopienas darbs ir ārkārtīgi nozīmīgs un varētu teikt – pat izšķirīgs korpusa veidošanā. Sabiedrības interesi var vērtēt atzinīgi, jo salīdzinoši īsā laikā visi pārrakstīšanai piedāvātie manuskripti arī tiek atšifrēti. Dienasgrāmatas un dzīvesstāsti folkloras digitālajā arhīvā ienes jaunu, personisku tematisko dimensiju, kas daudziem līdzstrādniekiem šķiet saistoša un uzmanības vērta. Veiksmīgajā sadarbībā ar sabiedrību liela nozīme ir LFK digitālā arhīva *garamantas.lv* iedibinātajām tradīcijām un digitālo līdzstrādnieku kopienai, kas kopš 2016. gada ar pieaugošām sekmēm palīdz atšifrēt folkloras manuskriptus (Reinsone 2018: 279–296; Reinsone 2020: 186–207).

Korpusa veidošana

Informācija par dienasgrāmatu korpusu veidošanu un digitālo metožu izmantošanu tā analīzē akadēmiskajā literatūrā ir visai skopa, un pētījumus par šo tēmu kopumā nav izdevies

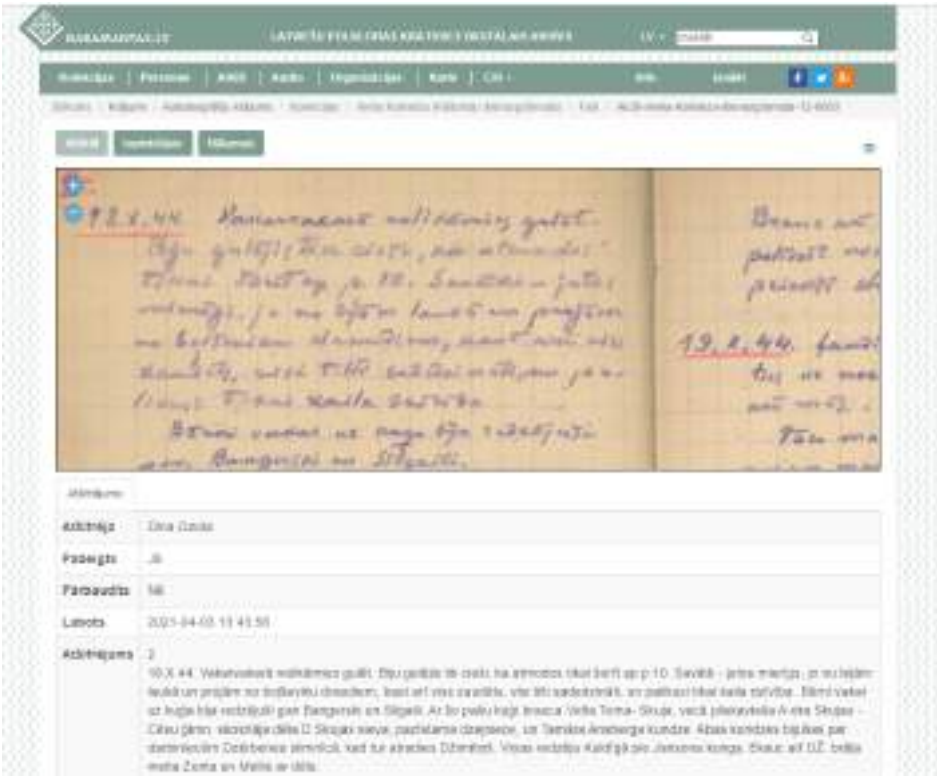
6 Piemēram, kolekcijā Nr. 12 ietilpst vien divas lapas no 12 gadu vecas meitenes dienasgrāmatas, kas rakstīta 1944. gadā (Jurča 1944).

7 Apjomīgākā dienasgrāmata Autobiogrāfiju krājumā ir Kaspara Aleksandra Irbes (1906–1996) no 1927. līdz 1996. gadam uz 8600 lappusēm sarakstīta dienasgrāmata (Lipša 2021: 415–442).

8 Pandēmijas dienasgrāmatu kolekcija tapusi, sākot ar 2020. gada martu: https://garamantas.lv/lv/collection/1415829/Pandemijas-dienasgramatas-2020.

9 Daļa materiālu, visvairāk dienasgrāmatas, nav publiski pieejami autortiesību, privātuma un citu iemeslu dēļ. Ar katru materiāla iesniedzēju ir noslēgta rakstiska vienošanās, kā šis materiāls tiek apstrādāts un kādas ir piekļuves tiesības.

10 Vairāk par Autobiogrāfiju krājumu sk. http://autobiografijas.lv.



2. attēls. Atšifrēšanas rīka saskarne *garamantas.lv*.

atrast pietiekamā skaitā. Pētnieciskajos rakstos uzsvars lielākoties likts uz dienasgrāmatu digitālu izdevumu izveidošanu. Piemēram, tiek reflektēts par dienasgrāmatu kodēšanu TEI XML formātā, anotējot atšifrējumā personas, vietas un citus konceptus (Thain 2016: 226–241). Tas digitālā izdevuma lasītājiem palīdz atklāt padziļinātu un bagātīgu kontekstu, kuru pētījuši un labi pārzina pētnieki, kas strādā pie šī projekta. Vēl kādā pētījumā datormetodes ir piemērotas (auto)biogrāfisko avotu pārveidošanai semantiski un lingvistiski anotētā korpusā, lai, izmantojot tīmekļa ontoloģijas valodu (*web ontology language*), varētu labāk izpētīt un interpretēt tekstā ietvertās zināšanu struktūras (Tóth 2013: 432–443). Savukārt kādā salīdzinoši nesenākā pētījumā ir izstrādāta jauna sistēma naratīvo pavedienu datorizētai identificēšanai dienasgrāmatām līdzīgiem tiešsaistes blogiem, izmantojot vairākus dabiskās valodas apstrādes paņēmienus (Bandeli et al. 2020: 63–69). Adelaidas Universitātes pētnieku grupa veidojusi I pasaules kara dienasgrāmatu korpusu, kur iekļautas vairāk nekā 500 dienasgrāmatas, kas aptver laiku no 1914. gada augusta līdz 1918. gada novembrim. Piemērojot vairākas tāllasījuma metodes, kā tēmu modelēšana (*topic modelling*) un sentimenta analīze, autori pētījumā

sniedz vispārinātu ieskatu, parādot tematiskās tendences un atsevišķu jēdzienu izplatību korpusā (Dennis-Henderson et al. 2020: 90–104). Līdz šim nav izdevies atrast pētījumus, kuros būtu analizēta vispusīga dienasgrāmatu tekstu korpusa veidošana vai arī kur šāds – vispārīgs un eventuāli neviendabīgs korpuss – būtu izmantots digitālo humanitāro zinātņu pētniecībā.

Pārlūkojot LFK Autobiogrāfiju krājumu, kas ir šī korpusa pamatā, redzam, ka iesniegtās un digitalizētās dienasgrāmatas ir stilistikas un formas ziņā samērā atšķirīgi materiāli, ko, protams, varēja paredzēt (Jackson 2010, 1–17, Lejeune 2009, 213–231). Autori kopumā dienasgrāmatās reflektējuši gandrīz visa 20. gadsimta garumā. Rakstītāji ir gan vīrieši, gan sievietes dažādos dzīves periodos, kā arī ir dienasgrāmatas, kuras rakstījuši vairāki autori, un to veikums nereti nav nošķirams. Dienasgrāmatām ir arī atšķirīga forma. Mēdz būt gan kalendāra veida dienasgrāmatas, kur līdztekus datumam fiksēts pavisam lakonisks dienas notikumu kopsavilkums, gan arī dienasgrāmatas, kur par ikdienu reflektēts daudz plašāk. Taču ir arī dienasgrāmatas, kurās šis nošķirums nav tik skaidrs – autors mēdz rakstīt pavisam strupi, tad kādā brīdī ieraksti kļūst garāki, citkārt atkal pievienojas arī cita veida teksti, piemēram, dzeja, lugu vai scenāriju fragmenti, lūgšanas, darba piezīmes vai lasīto grāmatu citēšana.

Atšķirīga ir arī valoda, ko atrodam krājuma materiālos. Dienasgrāmatās uzskatāmi redzam valodas attīstību gadsimta garumā – gan leksikas, gan sintakses ziņā. Tā kā LFK Autobiogrāfiju krājumā tiek iekļautas nepublicētas un nerediģētas dienasgrāmatas, tad dienasgrāmatās var sastapt nestandartizētu leksiku, netipiskas gramatiskās formas, senvārdus, izlokšņu vārdus, pierakstus no mūsdienu rakstu tradīcijas atšķirīgā ortogrāfijā, kā arī daudz gramatisko kļūdu un dažādu ortogrāfijas principu lietojumu. Dienasgrāmatas ir kultūrvēsturisks fenomens, kurš nav iegrožojams noteiktu formālu kritēriju standartos. Tas arī padara dienasgrāmatu korpusa izveidošanas procesu metodoloģisku sarežģījumu bagātu.

Lai dienasgrāmatu tekstu korpusu tā veidošanas sākumposmā padarītu cik iespējami prognozējamu un pārskatāmu, un tādu, kam varētu piemērot vienotu metodoloģisko analīzi, izmēģinājuma korpusā iekļauta tikai daļa no LFK Autobiogrāfiju krājumā iesniegtajām dienasgrāmatu kolekcijām. Proti, raksts koncentrējas uz tā sauktajām stāstošajām dienasgrāmatām, kur autori par savu ikdienu reflektē izvērstākā formā. Pieņemot, ka viens dienasgrāmatas ieraksts ir relatīvi patstāvīga vienība, garāka naratīvā refleksija par ikdienu var sniegt dziļāku ieskatu ne tikai autoru personībās, bet arī izklāstītajās tēmās un naratīva veidošanas paņēmienos. Tomēr jāpiebilst, ka, lai saglabātu autoru dzīves pierakstu integritāti, korpusā dienasgrāmatas tiek iekļautas pilnībā. Ja kādā no tām kādā brīdī parādās lakoniski ieraksti, arī tie tiek iekļauti korpusā.

2021. gada februārī latviešu dienasgrāmatu pilotkorpusu veidoja 270[11] dienasgrāmatas, kas kopā veidoja 17 560 000 rakstzīmju jeb 2 830 000 vārdlietojumu. Vidējais dienasgrāmatu ieraksta garums korpusā ir 1079 rakstzīmes. Pilotkorpuss, protams, aizvien nav viendabīgs. Atsevišķas dienasgrāmatas, kas rakstītas ļoti ilgu laiku jeb t. s. garās dienasgrāmatas,

11 Dienasgrāmatu skaits pārsniedz Autobiogrāfiju krājuma kolekciju skaitu, jo reizēm kolekcijās iekļauts vairāk par vienu dienasgrāmatu, piemēram, Pandēmijas dienasgrāmatu kolekcijā.

procentuāli aizņem lielu dienasgrāmatu korpusa daļu. Nenoteiktais un savstarpēji ļoti atšķirīgais apjoms ir dienasgrāmatu īpatnība, kas jāņem vērā, veidojot šādu korpusu. Līdzsvarotāka situācija varētu veidoties, ja tiek veidots dienasgrāmatu korpuss vienam laika posmam vai izmantotas noteikta veida dienasgrāmatas, piemēram, kara dienasgrāmata (Dennis-Henderson et al. 2020: 90–104).

Lai izveidoto pilotkorpusu varētu izmantot dienasgrāmatu tekstu diahroniskai analīzei un veikt dažāda veida salīdzinošos mērījumus starp autoriem, dzimumiem vai vecumgrupām, nepieciešams dienasgrāmatu tekstus saskaldīt atsevišķu dienu ierakstu failos, kas vēlāk ļautu tos kombinēt pēc dažādiem kritērijiem, piemēram, pēc desmitgadēm vai autoru vecumposma utt. Proti, tas ir kā veidot vienu kopējo dienasgrāmatu, kur visu autoru rakstītais būtu vienkop sakārtots hronoloģiskā secībā. Šī uzdevuma veikšanu sarežģī dažādais datumu pieraksts dienasgrāmatās[12].

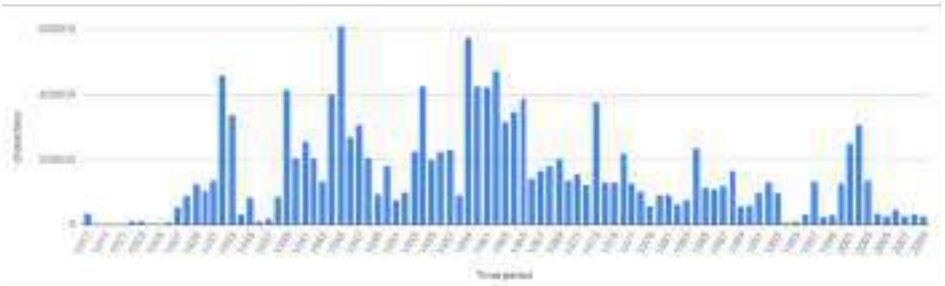
Veicot datumu pierakstu analīzi dienasgrāmatu korpusā ar mērķi noteikt datumu pieraksta sistēmu katrā atsevišķā dienasgrāmatā, atklājies, ka datumu pierakstīšana mēdz būt itin radošs process, kur liela nozīme ir autora gaumei, ieradumam un, iespējams, arī emocionālajam noskaņojumam. Datums var būt pierakstīts dienas ieraksta sākumā vai beigās, dažreiz arī pa vidu, bet citkārt ieraksts var būt arī bez datuma norādes, jo tas noprotams kontekstuāli no tekstā vēstītā, vai arī datuma vietā ir dota norāde uz svētkiem, kas tajā dienā svētīti. Ir autori, kuri gadskaitli min tikai gada sākumā, bet mēnesi norāda mēneša sākumā, dienas numurējot ar skaitļiem. Tiek izmantoti arī ļoti dažādi gada un mēnešu saīsinājumi, arī pieturzīmju (punkts, komats, slīpsvītra, kols, semikols, defise) lietojums ir daudzveidīgs, un visbeidzot mēdz būt arī kļūdaini norādīti mēneši, dienas vai gadi. Datumu ainava dienasgrāmatās ir patiešām daudzveidīgi krāšņa.

Piemēram, Dāvja Dauvarta dienasgrāmatā, kas rakstīta no 1932. līdz 1944. gadam, ir atrodami 15 dažādi datuma pieraksti.

31. 12. 31.	*3/I. 33.*	*2. jūlijā 1933. g.*	*7. septembra naktī*	*1933. g.* *10. Septemb.*	*1933. g.* *5. Septembrī*
š. g. 11. *septm.*	*33. g. 31. oktobrī*	*1933. gadā* *1. janvārī*	*1934. 2. janvārī*	*1935. g.*	*31/V. 36. g.*
12. aug. 1938. g	*6. jūlijā 1941. g.*	*1. 3. 44. g.*			

1. tabula. Datumu pieraksta veidi Dāvja Dauvarta dienasgrāmatā (1889–1944), *LFK Ak145* (Dauvarts 1932–1944).

12 Protams, būtu bijis vēlams, ka datumu vienādošana notiek tajā brīdī, kad dienasgrāmatas tiek atšifrētas, piemēram, izmantojot kalendāra rīku datumu anotēšanai (*calendar tool for date annotation*), taču pilotkorpusa veidošanas brīdī šāda iespēja vēl nebija izstrādāta.



3. attēls. Pilotkorpusā iekļauto dienasgrāmatu laika līnija pēc rakstzīmju skaita.

(Datu analīze veikta *Jupyter Notebook* lietojumprogrammā, izmantojot Python programmēšanas valodu, Pandas pakotni datu pārveidošanai un tīrīšanai, kā arī *Matplotlib* un *Plotly* pakotnes datu vizualizācijai. Daļa grafiku veidota, izmantojot *Google Sheets*.)

Datumu, kas norāda, kur sākas dienas ieraksts, varētu dēvēt par metadatumu. Tomēr datumi dienasgrāmatās izmantoti ne tikai kā metadatumi – gadi, mēneši, datumi, nedēļas, dienas un cita veida laika norādes (šodien, vakar, pērn utt.) ir bieži atrodamas arī pašā tekstā.

Tādējādi, lai izveidotu automatizētu rīku, kas atrastu visus metadatumus, tā ļaujot teksta dokumentu automātiski sadalīt pa dienu ierakstiem, bija veicami divi uzdevumi: (1) atrast pilnīgi visus datumus korpusā; (2) efektīvi atšķirt un nošķirt metadatumus no cita veida datumu pieminējumiem tekstā, kā arī no citām ciparu un burtu kombinācijām tekstā, kas izskatās līdzīgi datumu pierakstam, bet nav datumi.

Lai labāk izprastu dienasgrāmatu tekstos atrodamo metadatumu pierakstu, tos nosacīti varētu iedalīt divās atšķirīgās grupās, proti, absolūtie un relatīvie metadatumi. Absolūtā metadatuma pieraksts būtu, piemēram, 14.02.1957. vai 14.II.57. Savukārt relatīvie metadatumi ir izsecināmi kontekstuāli un no attiecīgā ieraksta pozīcijas kopējā dienasgrāmatas struktūrā. Tas nesagādā grūtības pētnieka veiktā tuvlasījumā, taču krietni apgrūtina automatizētu datuma noteikšanu. Piemēram, ja dienasgrāmatas tekstā attiecīgās dienas ieraksts sākas ar *14)*, tad automatizētajam datumu atpazīšanas rīkam ir jāspēj saprast vai atcerēties konteksts, ka vairākas lappuses iepriekš norādīts *februāris*, bet vēl iepriekš tekstā – *1957*. Tādējādi no kopējās teksta struktūras, datumu secības un konteksta ir secināms, ka attiecīgais metadatums 14) apzīmē 1957. gada 14. februāri.

Visbiežāk tomēr metadatumu izvietojums dienasgrāmatu tekstos ir šāds: tukša rinda pirms jaunas dienas ieraksta, metadatums – jaunā rindā, attiecīgās dienas ieraksts – jaunā rindā. Lai pareizi veiktu atsevišķo dienasgrāmatu priekšapstrādi teksta sadalīšanai atsevišķu dienu ierakstos, būtiski ņemt vērā katrā izmantoto datēšanas sistēmu un attiecīgi pielāgot metadatumu atpazīšanas rīku. Taču kopumā rezultāts, veicot izlases veida manuālu pārbaudi, bija teicams – korpuss bija pareizi

4. attēls. Vienas dienas ieraksts datubāzē.

sadalīts pa dienas ierakstiem. Kopumā pilotkorpusu veido 16 904 dienas pieraksti, kas aptver laika posmu no 1917. līdz 2021. gadam.

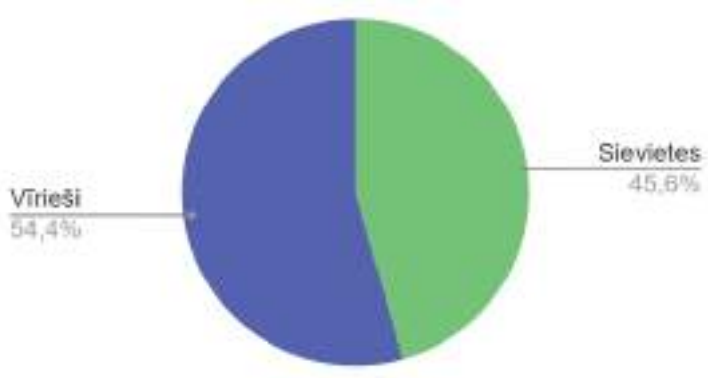
Iegūstot šādas – pa dienas ierakstiem sadalītas – dienasgrāmatas, paveras jaunas iespējas analizēt to, kā šis pēc nejaušības principa veidotais tekstu korpuss izskatās dažādos šķērsgriezumos. 3. attēlā redzams, kā dienasgrāmatu ieraksti pēc zīmju skaita izkārtojas cauri gadiem. Korpusā iekļautie teksti dokumentē laiku no 1917. gada līdz pat mūsdienām. Kaut arī dienasgrāmatu pierakstos nosegts viss redzamais laikposms, tomēr dokumentējums nav viendabīgs. Redzams, ka izteikti vairāk ir dienasgrāmatu ierakstu, kas tapuši ap Otrā pasaules kara beigām, un izceļas arī 20. gadsimta 60. gadu sākuma posms. Šajā grafikā gan nav integrētas 2020. un 2021. gadā pandēmijas laikā tapušās 239 dienasgrāmatas, jo to īpatsvars hronoloģiskā griezumā ir daudzkārt lielāks nekā pārējām dienasgrāmatām, kas šāda veida grafiku padarītu grūtāk uztveramu.

Sadalot dienasgrāmatu tekstu korpusu pa dienas ierakstiem, katram no tiem tiek piešķirti metadati, kuri norāda: (1) dienasgrāmatas kolekcijas numuru Autobiogrāfiju krājumā, kas ļauj atlasīt visus viena autora ierakstus; (2) ieraksta numuru pēc kārtas dienasgrāmatā, kas neļauj kļūdīties ar ierakstu secību, ja vairāki ieraksti tapuši vienā dienā; (3) metadatumu, kas norāda uz ieraksta tapšanas laiku – gan autora oriģinālajā pieraksta formā, gan pārveidots vienotā *sql* datubāzes datuma pierakstā; (4) dienasgrāmatas ieraksta rakstzīmju skaitu. Papildus tam ir izveidotas relācijas ar autoru metadatiem, kas savukārt ļauj veidot dienasgrāmatas ierakstu izlases jeb apakškorpusus pēc dažādiem kritērijiem, piemēram, pēc autora dzimumu vai/un vecuma posma, kā arī pēc noteikta vecuma posma noteiktā laika periodā.

Vīriešu un sieviešu dienasgrāmatas

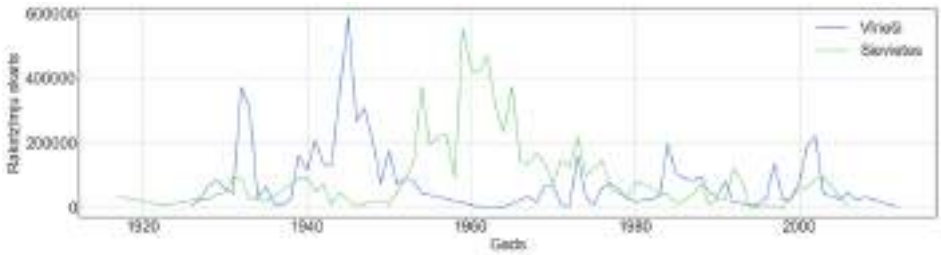
Dzimumu reprezentācija dienasgrāmatu korpusā varētu būt viens no svarīgākajiem jautājumiem, jo īpaši, lai noskaidrotu, vai ir iespējams izveidot divus līdzvērtīgus tekstu apakškorpusus – sieviešu un vīriešu –, lai tos salīdzinoši pētītu, izmantojot tāllasījuma metodes. Turklāt, tā kā dienasgrāmatu tekstu korpuss ir piesaistīts laikam, ir svarīgi arī noskaidrot, kā abi dzimumi ir pārstāvēti dažādos laika posmos. Mēs apzināmies, ka dzimumu jautājums nav viennozīmīgs un vienkāršs un ka eventuāli varētu tikt veidoti arī citi dalījumi apakškorpusos, ņemot vērā dienasgrāmatu autoru pašidentifikāciju. Šajā korpusā balstīto nākotnes pētījumu mērķis varētu būt noskaidrot, vai formālā ziņā ir pamats nošķirt sieviešu un vīriešu dienasgrāmatas. Varbūt tematiskā, emocionālā un stilistiskā līdzība nav dzimuma determinēta, bet būtisks ir laika periods, vecuma posms vai kāds cits kritērijs?

Analizējot korpusa metadatus dzimumu perspektīvā, no tā tika izņemtas ārā tās dienasgrāmatas, kuras ir rakstījuši kolektīvi dažādu dzimumu pārstāvji, un arī tās, kuru autoru dzimums nav zināms.



5. attēls. Sieviešu un vīriešu rakstījumu reprezentācija korpusā pēc rakstzīmju skaita.

Pārlūkojot korpusā iekļautos tekstus vispārīgi, ir redzama diezgan līdzsvarota aina. Vīriešu rakstītais tekstuālais apjoms ir tikai nedaudz lielāks nekā sievietēm (sk. 5. attēlu). Tomēr šāds vienkāršots skatījums neļauj apjaust to, kā korpuss ir veidojies dzimumu perspektīvā attiecībā pret laiku. Proti, cik līdzsvaroti vīriešu un sieviešu rakstītais ir pārstāvēts dažādos laika posmos. Vai varam runāt par kādu periodu, kurā pārliecinoši naratīvā dominē viens vai otrs dzimums? Ņemot vērā to, ka dienasgrāmatas nav standartizētas publikācijas, kuru izdošanas biežumu un apjomu varētu iepriekš prognozēt, tad detalizētāku ainu par to, kā izraudzīto dienasgrāmatu ieraksti izskatās laika līnijā, mēs ieraugām tikai tad, kad ir veikta padziļinātāka materiālu apstrāde un analīze.



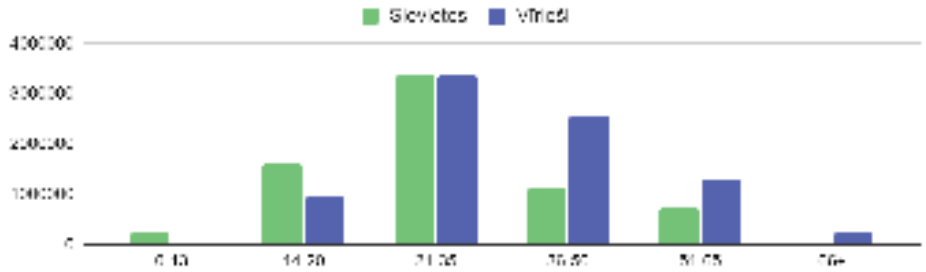
6. attēls. Sieviešu un vīriešu rakstīšanas intensitāte autobiogrāfisko tekstu korpusā.

6. attēls vizuāli attēlo to, kā latviešu dienasgrāmatu tekstu pilotkorpusā ir reprezentētas vīriešu un sieviešu dienasgrāmatas noteiktos laika periodos. Kaut arī, veidojot korpusu un strādājot ar kolekcijām, šķita, ka situācija šajā ziņā varētu būt samērā līdzsvarota, tomēr metadatu analīze parāda interesantu ainu. Uzskatāmi redzams, ka laiks no 1940. līdz pat 50. gadu sākumam dienasgrāmatās dokumentēts no vīriešu perspektīvas. Jāpiebilst, ka krājumā un arī tekstu korpusā nav mērķtiecīgi meklētas un iekļautas karavīru dienasgrāmatas, tādas atrodamas tikai dažas (Sīlis 1943–1946; Upmalis 1944–1946).

Sieviešu naratīvās balsis sāk dominēt 50. gadu sākumā, un to dominance turpinās divas desmitgades. Vēlāk situācija kļūst līdzsvarotāka, bet būtiski samazinās arī pilotkorpusa tekstu apjoms. Šāda izteikti viena dzimuma dominēšana ilgākā laika periodā kopējai korpusa sabalansētībai ir nevēlama, un jebkuri vispārinoši secinājumi par korpusu kopumā vai noteiktā laika posmā ir jāizdara ar piesardzību un paturot prātā tā formālās īpatnības. Taču no pagaidu korpusa padziļinātākai pētniecībai pamatoti varētu izvēlēties vīriešu dienasgrāmatas, kas tapušas ap Otrā pasaules kara laiku, un sieviešu dienasgrāmatas padomju okupācijas pirmajās desmitgadēs.

Vecumgrupu rakstījumu pārstāvniecība pilotkorpusā

Viens no vērtīgākajiem ieguvumiem korpusa metadatu analīzē ir iespēja izveidot relāciju starp autora vecumu un rakstīšanas vēsturisko laiku. Respektīvi, ja 1960. gadā autoram ir 10 gadu, tad viņš pārstāv vienu vecuma grupu. Bet, ja šim pašam autoram 1982. gadā ir 32 gadi, tad viņa rakstītais jau ir iekļauts citā vecuma grupā. Šāda informācija ļauj pārlūkot korpusu tādā perspektīvā, kādu bez datormetožu izmantojuma būtu ārkārtīgi sarežģīti īstenot. Balstoties uz šādu analīzi, ir iespējams pamatotāk izraudzīties tekstu analīzei piemērotākās vecuma grupas konkrētos laika periodos.



7. attēls. Sieviešu un vīriešu rakstījumu apjoma salīdzinājums pēc vecumgrupām.

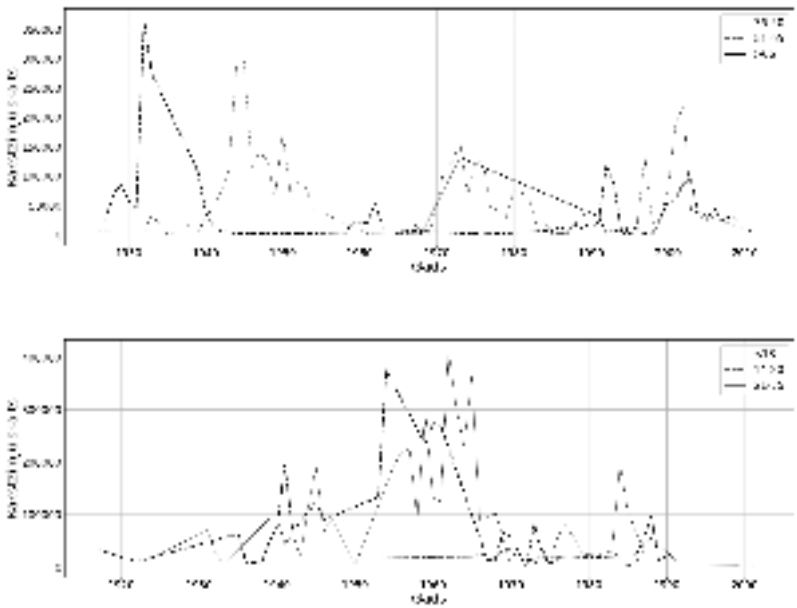
Dienasgrāmatu tekstu pilotkorpusā sākotnēji izraudzīts sadalīt dienasgrāmatu ierakstus pa desmitgadēm. Proti, jauna vecuma grupa tiek veidota ikkatrus desmit gadus (< 10 g. v., no 11 līdz 20 g. v. utt.). Pārlūkojot korpusu šādā vecumgrupu sadalījumā, tika secināts, ka šī pieeja nav pamatota. Liela daļa vecumgrupu pilotkorpusā bija ļoti vāji reprezentētas, un kopaina bija pārāk sadrumstalota, lai sniegtu jelkādu papildu informāciju. Ņemot to vērā, dienasgrāmatu tekstu pilotkorpusā tika noteiktas sešas vecuma grupas, par pamatu ņemot vispārīgus sociālos, izglītības un ekonomiskos aspektus, kas nosaka cilvēka dzīves posmus. Pirmā grupa: bērni līdz 13 gadu vecumam. Otrā grupa: jaunieši no 14 līdz 20 gadu vecumam. Trešā grupa: jaunie pieaugušie no 21 līdz 35 gadu vecumam. Ceturtā grupa: pieaugušie I no 35 līdz 50 gadu vecumam. Piektā grupa: pieaugušie II no 51 līdz 65 gadu vecumam. Sestā grupa: seniori virs 66 gadu vecuma. Protams, arī šāds iedalījums nav viennozīmīgs, jo patstāvīgas dzīves sākums dažādos laikmetos varētu būt atšķirīgs. Veicot tekstu analīzi, iespējams, vecuma grupas nāksies koriģēt.

Pārlūkojot šādu vecumgrupu sadalījumu dzimumu perspektīvā, redzam, ka bērnu un senioru vecumgrupa ir ārkārtīgi vāji pārstāvēta korpusā un būtu apsverams, vai šīs abas grupas nebūtu lietderīgi apvienot attiecīgi ar 3. grupu (jaunieši) un 5. grupu (pieaugušie II). Līdzsvarota situācija dzimumu perspektīvā vērojama 21–35 gadu vecu cilvēku vecumgrupā, savukārt pārējās dominē viens vai otrs dzimums.

Attēlojot vecumgrupu rakstījumu intensitāti uz vienas laika līnijas (7. attēls), iegūstam papildu informāciju par to, kas jāņem vērā, veicot turpmāku analīzi.

8. un 9. attēls parāda, ka bērnu dienasgrāmatas korpusā ir izkaisītas cauri gadsimtam vienmērīgi mazā apjomā. Taču citu vecumgrupu reprezentācija uz laika līnijas iezīmē noteiktus periodus, kuros izpēte varētu būt sekmīgāka, jo pieejams vairāk materiāla. Piemēram, jauniešu grupas (14–20 g. v.) dienasgrāmatu ieraksti lielā intensitātē tapuši II pasaules kara laikā, bet visvairāk no 1955. līdz 1965. gadam.

Jauno pieaugušo (21–35 g. v.) vecumgrupas intensīvākais rakstīšanas laiks korpusā ir no 50. gadu vidus līdz 60. gadu beigām, ar atsevišķiem intensitātes punktiem ap Otrā pasaules kara beigām un 80. gadu vidū. Turklāt, atsaucoties uz 6. attēlu, šī ir arī grupa, kur abi



8. un 9. attēls. Korpusa sadalījums laikā pēc vecuma grupām.

dzimumi ir vienlīdz labi pārstāvēti. Pieaugušo grupa I (36–50 g. v.), kur dominē vīriešu naratīvās balsis, vislabāk reprezentēta ap Otrā pasaules kara laiku līdz pat 50. gadu sākumam. Otrs intensīvākais rakstīšanas posms šai vecuma grupai ir 70. gadi un arī – ap tūkstošgades miju. Pieaugušo grupai II (51–65 g. v.) ir izteikta rakstīšanas intensitāte 30. gados, kā arī ap 1990. gadu, savukārt pārējās desmitgadēs tā ir nemanāma. Savukārt senioru grupa ir maz reprezentēta līdz pat 60. gadu beigām. Pārējo vecumgrupu kontekstā senioru rakstītais veido būtisku pilotkorpusa daļu no 1970. līdz aptuveni 1990. gadam, kas varētu liecināt, ka šīs grupas apvienošana ar pieaugušo grupu II varētu nebūt pamatota.

Atgriežoties pie 3. attēla, kur labi redzams, ka 2. pasaules kara beigas ir visintensīvāk dokumentētais laikmets dienasgrāmatu pilotkorpusā, šajos – 8. un 9. attēlā – redzam, ka autori, kas rakstījuši ap 1945. gadu, pārstāv trīs vecuma grupas – jauniešus, jaunos pieaugušos un pieaugušo I grupu. Un, salīdzinot šos ar 6. attēlu, redzam, ka šajā laikā dominē vīriešu dzīves pieraksti. Respektīvi, padziļināti analizējot šo laika posmu, būtu jāņem vērā, ka II pasaules kara laiku reprezentē salīdzinoši jaunu vīriešu naratīvās balsis.

Lai gan aina joprojām izskatās neviendabīga un sadrumstalota, jāņem vērā, ka metadati tiek analizēti trijās dimensijās – dzimums, vecumposms un laikmets –, kas katra atsevišķi saskalda sīkāk aplūkojamo korpusa materiālu. Tomēr šāda analīze labi parāda katras vecuma grupas

pārstāvību korpusā dažādos laika posmos. Tas ir noderīgi, lai izvēlētos nākamos soļus gan tāllasījuma, gan kvalitatīvās analīzes veikšanai. Papildus būtu vērts arī sīkāk izpētīt, kā veidojas šie intensitātes periodi, t. i., cik lielu īpatsvaru grupas kopējā pārstāvībā sniedz atsevišķu autoru rakstījumi. Jāņem vērā, ka intensīvi rakstošam autoram varētu būt ievērojama ietekme uz kādas vecuma grupas pārstāvniecību attiecīgajā periodā.

Secinājumi

Metadatu analīzē balstīti mērījumi, kuri veikti, pētot dienasgrāmatu tekstu pilotkorpusu, ir izmantojami secinājumiem par korpusu vai kolekciju, tie var iezīmēt eventuālas tendences vispārējos dienasgrāmatu rakstīšanas paradumos, tomēr ar šiem datiem vien nepietiek, lai izdarītu pamatotus secinājumus par dienasgrāmatu rakstīšanas tradīciju un tendencēm Latvijā.

Vispārēja dienasgrāmatu tekstu korpusa veidošana kopumā ir sarežģīts projekts. Ideālā gadījumā varētu vēlēties redzēt vienlīdz labi pārstāvētus abus dzimumus un vienlīdz labi pārstāvētus abus dzimumus visās vecuma grupās. Mēs varētu arī vēlēties vienlīdz labi pārstāvētas visas vecuma grupas kopumā dažādos laika posmos. Tomēr jāatminas dienasgrāmatu tekstu korpusa veidošanas īpatnības. Proti, dienasgrāmatu apjoms, forma, veids un arī autoru dzimumpiederība vai vecums nav prognozējams un ietekmējams. Veidojot kolekciju un attiecīgi arī korpusu, ir iespēja mērķtiecīgāk meklēt dienasgrāmatas pēc noteiktiem kritērijiem, taču lielākoties izvēle nav liela un nākas rēķināties ar to, kas ir pieejams arhīvos, bibliotēkās vai personīgajās kolekcijās. Šis raksts pievēršas vien autoru dzimumam un vecumam, nediskutējot par izglītības līmeni, dzīvesvietu vai sociālo stāvokli. Bieži vien šāda informācija par dienasgrāmatu autoriem nav zināma, taču iespējamo kritēriju skaits, pēc kādiem sadalīt dienasgrāmatas apakškorpusos un kādā perspektīvā pētīt, var būt krietni liels.

Raugoties no kultūvēsturiskās perspektīvas, būtu nepieciešams saglabāt un datorizēt visas pieejamās dienasgrāmatas. Savukārt turpmākajā dienasgrāmatu korpusa attīstībā drīzāk būtu jāturpina fokusēties uz eventuāli apjomīgākajiem apakškorpusiem un arī tiem apakškorpusiem, kuri pētniekiem ir nepieciešami konkrētu pētniecisko tēmu izzināšanai.

Kaut arī dienasgrāmatu tekstu pilotkorpusa apjoms šobrīd nav mazs, tomēr, sastatot tekstus attiecībā pret laiku (gadu, desmitgadi), kurā fiksēts dienasgrāmatas ieraksts, laikmeta pārklājums ir samērā plāns. Ir gadi, kurus pārstāv līdz pat desmit autoriem, bet ir periodi, kad rakstītāju skaits ir visai niecīgs. Autoru skaita palielināšana varētu būt vēl viens dienasgrāmatu teksta korpusa nākotnes uzdevums. Tiek plānots iekļaut pilnīgi visas digitalizētās, ne tikai stāstošās dienasgrāmatas, metadatos papildus norādot dienas ieraksta garumu, lai turpmākajos izmēģinājumos ar teksta analīzi ierakstus ar noteiktu apjoma ierobežojumu varētu definēt pēc nepieciešamības.

Dauvarts, Dāvis (1932–1944). Dāvja Dauvarta dienasgrāmata. LFK AK 145. Pieejams: http://garamantas.lv/lv/collection/1362666/Davja-Dauvarta-dienasgramata [skatīts 15.10.2021.].

Dennis-Henderson, Ashley, Roughan, Matthew, Mitchell, Lewis, Tuke, Jonathan (2020). Life Still Goes on: Analysing Australian WW1 Diaries through Distant Reading. *Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. International Committee on Computational Linguistics, pp. 90–104.

Hockey, Susan (2004). The history of humanities computing. Schreibman, Susan, Siemens, Raya Unsworth, John (eds). *A Companion to Digital Humanities*. Oxford: Blackwell, pp. 1–19.

Jackson, Anna (2010). *Diary Poetics. Form and Style in Writers' Diaries, 1915–1962*. New York: Routledge.

Jurča, Ērika (1944). Ērikas Jurčas dienasgrāmata. LFK AK 15. Pieejams: http://garamantas.lv/lv/collection/1134078/Erikas-Jurcas-dienasgramata [skatīts 15.10.2021.].

Bandeli, Kumar Kiran, Hussain, Muhammed Nihal, Agarwal, Nitin. (2020). A Framework towards Computational Narrative Analysis on Blogs. *CEUR Workshop Proceedings,* vol. 2593, pp. 63–69.

Koncar, Philipp, Fuchs, Alexandra, Hobisch, Elisabeth, Geiger, Bernhard, Scholger, Martina, Helic, Denis (2020). Text sentiment in the Age of Enlightenment: an analysis of spectator periodicals. *Applied Network Science*, No. 5(1).

Lejeune, Philippe (2009). *On Diary*. Honolulu: University of Hawaii Press.

Lipša, Ineta (2021). Documenting the queer self: Kaspars Aleksandrs Irbe (1906–1996) in between unofficial sexual knowledge and medical-legal regulation in Soviet Latvia. *Cahiers du monde russe*, No. 62/2-3, pp. 415–442.

Moretti, Franco (2013). *Distant Reading*. London: Verso.

Nyhan, Julianne, Flinn, Andrew (2016). *Computation and the Humanities: towards an Oral History of Digital Humanities*. Springer Open.

Rak, Julie (2019). Big Data and Self-Tracking: Research Trajectories. Barnwell, Ashley, Douglas, Kate (eds.). *Research Methodologies for Auto/biography Studies*. London: Routledge, pp. 116–121.

Reinsone, Sanita (2018). Participatory practices and tradition archives. Harvilahti, Lauri, Kjus Audun, Cliona, O'Caroll, Österlund-Pötzsch, Sussane, Skott, Frederik, Treija, Rita (eds.). *Visions and Traditions: Knowledge Production and Tradition Archives*. Helsinki: SKS Academia Scientiarum Fennica, pp. 279–296.

Reinsone, Sanita (2020). Searching for deeper meanings in cultural heritage crowdsourcing. Hetland Per, Pierroux, Palmyre, Esborg, Line (eds.). *A History of Participation in Museums and Archives. Traversing Citizen Science and Citizen Humanities*. London: Routledge, pp. 186–207.

Sula, Chris Alen, Hill, Heather (2019). The early history of digital humanities: An analysis of Computers and the Humanities (1966–2004) and Literary and Linguistic Computing (1986–2004). *Digital Scholarship in the Humanities*, No. 34, pp. 190–206.

Sīlis, Kārlis Alberts (1943–1946). Kārļa Alberta Sīļa dienasgrāmata. LFK AK 174. Pieejams: http://garamantas.lv/lv/collection/1458840/Karla-Alberta-Sila-dienasgramata [skatīts 16.10.2021.].

Thain, Marion (2016). Perspective: Digitizing the Diary – Experiments in Queer Encoding (A Retrospective and a Prospective). *Journal of Victorian Culture*, No. 21 (2), pp. 226–241.

Tóth, Gábor Mihály (2013). The computer-assisted analysis of a medieval commonplace book and diary (MS Zibaldone quaresimale by giovanni rucellai). *Literary and Linguistic Computing*, No. 28 (3), pp. 432–443.

Upmalis, Laimonis (1944–1946). Laimoņa Upmaļa dienasgrāmata. LFK AK 148. Pieejams: garamantas.lv/lv/collection/Laimona-Upmala-dienasgramata [skatīts 16.10.2021.].

Summary

Metadata-based Analysis of the Diaries' Corpus

Sanita Reinsone, Haralds Matulis,
Ilze Ļaksa-Timinska

Keywords: digital humanities, distant reading, text analysis,
diary corpus, autobiographical texts

Personal diaries are a unique narrative document of an era, revealing the times and events through a deeply personal perspective. Each diary is undoubtedly unique, having been written in a particular place and time, encapsulating many of the authors' present-day experiences. However, together they create a special dimension of personal writing, where the individual style is shaped by the author's taste and writing skills, as well as by the influences of diaristic traditions, fiction and other cultural phenomena. Recognizing the uniqueness and particularity of each diary, in this paper we will focus on personal diaries as a whole and conduct a metadata-driven analysis of a pilot corpus of diaries with the aim of revealing the particularities of the corpus in the juxtaposition of different criteria.

This paper is based on recent experience of building a corpus of personal diary texts from the very beginning, i.e. starting with the collection, digitisation and further processing of the material to build the corpus. In this article, we will analyse the creation of the corpus of diary texts, the methodological challenges and the difficulties encountered in such an initiative.

The metadata-based measurements that we have taken in the pilot corpus of diary texts can be used to draw conclusions about a corpus or a collection, they can highlight possible trends in general diary-writing habits, but these data alone are not sufficient to draw valid conclusions about the tradition and trends of diary-writing in Latvia.