10.35539/LTNC.2022.0047.A.B.V.S.150.167

Latento Dirihlē sadalījumu modeļa izmantojums laikraksta *Latvijas Kareivis* tematu analīzē: Oskara Kalpaka gadījuma izpēte

Anda Baklāne, Valdis Saulespurēns

Publikācija tapusi Valsts pētījuma programmas "Humanitāro zinātņu digitālie resursi" projektā "Humanitāro zinātņu digitālie resursi: integrācija un attīstība" (Nr. VPP-IZM-DH-2020/1-0001).

Atslēgvārdi: tematu modelēšana, digitalizētie laikraksti, digitālā vēsture, tematu koherence, Latvijas Nacionālā bibliotēka

Ievads

Jau kopš 1999. gada Latvijas Nacionālā bibliotēka (LNB) veic vēsturisko laikrakstu, grāmatu, attēlu, audio un video kolekciju digitalizāciju (Krūmiņa 2012; Zariņš 2014). Teksta kolekcijām tikusi pievērsta vislielākā vērība; tiek lēsts, ka digitalizēto laikrakstu kolekcijas ietver vairāk nekā 80 % periodikas materiālu, kas publicēti līdz 20. gadsimta 90. gadu vidum[1]. Materiāli tikuši segmentēti un optiski atpazīti, tādējādi to lietotāji var izmantot iespējas, kuras sniedz iespēja meklēt vārdus pilnajā tekstā. Tomēr, sekojot valodas tehnoloģiju attīstībai un pašreizējām tendencēm digitālo humanitāro zinātņu izpētē, pastāv pieprasījums pēc jaunu pakalpojumu izveides, kas sniegtu vēl vairāk iespēju padziļinātai teksta dokumentu izpētei (Ehrmann et al. 2020; Ūdre et al. 2019).

Šajā rakstā gadījuma izpēte veikta ar mērķi pētīt un pārbaudīt, kā automātiskas tematu noteikšanas metodes varētu tikt izmantotas LNB digitalizēto periodisko izdevumu kolekciju pārlūkošanā un izpētē. Tematu noteikšana (*topic detection*) jeb tematu modelēšana (*topic modelling*) bijusi viena no biežāk izmantotajām teksta digitālas analīzes tehnikām sociālajās un humanitārajās zinātnēs 21. gadsimta pirmajās desmitgadēs, savukārt latento Dirihlē sadalījumu (*latent Dirichlet allocation*, LDA) metodoloģija – viena no biežāk izmantotajām tematu noteikšanas metodēm; tātad LDA jau ir labi izprasta un pārbaudīta dažādos lietojumgadījumos.

LDA izmantojums vēsturisko laikrakstu kolekciju pētniecībai latviešu valodā tomēr uzskatāms par novatorisku. Dabīgās valodas apstrādes tehnoloģijas visstraujāk attīstījušās lielajās valodās, savukārt mazās valodās ar sliktu valodas tehnoloģiju un resursu nodrošinājumu jauninājumi tiek ieviesti novēloti (Abney 2010). Saskaņā ar META-NET aplēsi 2012. gadā latviešu valodas atbalsts teksta un valodas apstrādes uzdevumu veikšanai tika vērtēts kā nepilnīgs vai neesošs (Skadiņa et al. 2012). Laikā, kad top šī publikācija, ir pieejami daudzi no dabīgās valodas apstrādes resursiem un rīkiem, kas 2012. gadā vēl nebija izstrādāti, to vidū apjomīgi anotēti korpusi[2], latviešu valodas teksta automātiskas morfoloģiskas un sintaktiskas marķēšanas rīki[3], tiek veidota latviešu valodas leksēmu semantisko

1 Latvijas Nacionālās digitālās bibliotēkas Periodikas portāls. Pieejams: http://www.periodika.lv/ [skatīts 18.06.2022.].

2 Latviešu valodas teksta un runas korpusi. Pieejams: korpuss.lv [skatīts 18.06.2022.].

3 NLP-PIPE: Latvian NLP Pipeline as a Service. Pieejams: http://nlp.ailab.lv/ [skatīts 18.06.2022.].

attiecību datubāze *WordNet*[4]. Tomēr, attīstoties jaunām tehnoloģijām, piemēram, vārdlietojumu kartēšanai (*word embeddings*) un īpaši lielu valodas modeļu veidošanai, valodas, kurām vēsturiski bijis zems tehnoloģiju un valodas resursu nodrošinājums, joprojām atpaliek (Alabi et al. 2020). 24 oficiālo Eiropas Savienības valodu vidū 15 valodas (t. sk. latviešu valoda) var tikt uzskatītas par valodām ar nepietiekamu resursu nodrošinājumu (Alves et al. 2020). Šajā publikācijā aplūkotais pētījums ir pirmais LDA metodoloģijas izmantojums latviešu vēsturisko laikrakstu analīzē[5] un kopumā viens no pirmajiem LDA izmantojumiem, analizējot tekstu latviešu valodā[6].

Dabīgās valodas apstrādes tehnoloģiju ieviešana vēsturisko dokumentu kolekcijās saistīta ar dažādiem šķēršļiem, ar kuriem nenākas saskarties, analizējot digitāli radītas, liela apjoma, stilistiski viendabīgas un standartizētas teksta datu kopas. Digitalizētu senu tekstu kopas nereti ietver relatīvi mazus korpusus, kas nav viendabīgi un līdzsvaroti, tādējādi ierobežojot mašīnmācīšanās metožu izmantošanas iespējas (McGillivray 2021). Laika posms no 19. gs. otrās puses līdz 20. gs. sākumam jau piedāvā salīdzinoši bagātīgu teksta dokumentu klāstu, taču digitālo analīzi apgrūtina dažādi vārdu pieraksta varianti, ar ortogrāfijas reformām saistītās rakstības izmaiņas un optiskās atpazīšanas kļūdas (Bollmann 2019). Oskara Kalpaka gadījuma izpēte apliecina, ka šī problemātika aktuāla, arī analizējot digitalizēto laikrakstu kolekcijas latviešu valodā.

Rakstā sadaļā "LDA metodoloģija humanitāro zinātņu pētniecībā" skaidrots LDA jēdziens, aplūkotas atšķirīgas pieejas LDA rezultātu interpretācijā un iezīmēti vairāki scenāriji, kā tematu modeļus varētu integrēt digitalizēto laikrakstu kolekciju saskarnēs. Sadaļā "Datu kopa" atrodama informācija par izmēģinājumā izmantotās datu kopas parametriem un datu atlases pamatojums. Sadaļā "Gadījuma izpētes rezultāti" izklāstīti un interpretēti datu modeļa apmācības rezultāti. Secinājumu daļā apkopotas pētījuma gaitā gūtās atziņas un norādīti iespējamie turpmākie soļi.

LDA metodoloģija humanitāro zinātņu pētniecībā

Tematu modelēšanas metodoloģijas ietver dažādas dabīgās valodas apstrādes un mašīnmācīšanās tehnikas, kas tiek izmantotas, lai atrastu liela apjoma dokumentu kolekcijās ietvertas satura struktūras jeb tematus. Tematu modelis ir matemātiskas varbūtības sadalījums no kādas dokumentu kolekcijas izgūtā vārdu kopā, kas paredz, kādiem tematiem pieder šie vārdi un kuri temati ir klātesoši katrā dokumentā.

4 Latvian WordNet. Pieejams: https://wordnet.ailab.lv/ [skatīts 18.06.2022.].

5 Provizoriskie Oskara Kalpaka gadījuma izpētes rezultāti aplūkoti rakstā: Baklāne, Saulespurēns 2022.

6 R. Vīksna, M. Kirikova un D. Kiopa veikuši Latvijas tiesību aktu tematu analīzi (kopumā vairāk nekā 50 000 dokumentu). Pētījumā salīdzinātas trīs tematu analīzes metodes, t.sk. LDA (Vīksna et al. 2020).

Tematu modeli lietotājs parasti var aplūkot kā vārdu sarakstus, kuros katram vārdam piešķirta atbilstošā varbūtības vērtība. Tiek sagaidīts, ka šiem vārdu sarakstiem jābūt semantiski saskanīgiem, taču ne visi tekstu žanri un ne visas dokumentu kolekcijas var kalpot par pamatu tematiski vienotu vārdu sarakstu radīšanai. Piemēram, mākslinieciski teksti vai neviendabīgas kolekcijas, kurās sajaukti dažāda garuma un dažādu žanru teksti, var nenodrošināt pamatu saskanīgu un viegli interpretējamu atslēgvārdu sarakstu veidošanai. Tomēr arī neviennozīmīgi interpretējami vārdu saraksti var tikt izmantoti korpusu izpētes vajadzībām, t. i., tematu modelēšanas algoritmu darba rezultāti var tikt izmantoti, ne vien lai lakoniski un nepārprotami atspoguļotu saturu, bet arī plašāk – lai vizualizētu, pētītu, izvirzītu hipotēzes par korpusu (Blei 2012).

Latento Dirihlē sadalījumu metode tematu atrašanai pirmo reizi aplūkota publikācijā 2003. gadā (Blei et at. 2003) un pašlaik tiek uzskatīta par vienu no biežāk izmantotajām tematu modelēšanas tehnikām (Marjanen et al. 2020; Pääkkönen, Ylikoski 2020). LDA ir varbūtības modelis, kas balstīts divos pieņēmumos: (1) pastāv noteikts skaits dotajos dokumentos bieži kopā lietotu vārdu kopumu (tematu); (2) katrs dokuments korpusā var saturēt vairākus tematus dažādās nozīmīguma pakāpēs (Blei 2012). Formāli LDA var definēt kā varbūtības $P(\theta 1{:}M, z1{:}M, \beta 1{:}k \mid D; \alpha 1{:}M, \eta 1{:}k)$ atrašanu. Tas nozīmē kopīgas varbūtības atrašanu M dokumentiem ar šādiem nezināmajiem: θ – tematu sadalījums pa vienam uz katru dokumentu; z – temati katrā dokumentā; β – vārdu sadalījums katrā tematā, k – kopējais tematu skaits visos dokumentos. Kā kopēja dotā varbūtība ir dots korpuss D un parametri: α – parametru vektors katram dokumentam, η – parametru vektors katram tematam. Šādu varbūtību nevar atrisināt ar standarta aritmētiskām metodēm (Blei et al. 2003). LDA atrisina šo varbūtību ar iteratīvu algoritmu, kas turpina modeļa apmācību, līdz tiek sasniegts zināms konverģences līmenis.

Varbūtiskās tematu modelēšanas lietojums var atšķirties atkarībā no pētniecības jomas un pētnieka mērķiem. Kopš 2003. gada tikuši veikti daudzi pētījumi, kuros LDA metodoloģija un tās atvasinājumi izmantoti, lai veidotu tematu modeļus zinātnisku publikāciju kolekcijām (Blei, Lafferty 2007; Newman et al. 2006; Hall et al. 2008), kā arī lai pētītu vēsturiskos laikrakstus un žurnālus (Block 2006; Nelson 2011; Templeton et al. 2011; Hengchen 2017). Spriežot pēc piemēriem, kurus piedāvā varbūtisko tematu modelēšanas metožu izstrādātāji, metode primāri tikusi veidota lietišķu teksta žanru, nevis daiļliteratūras izpētei, taču laika gaitā tikusi analizēta arī mākslinieciskā proza un dzeja (Rhody 2012).

Koherentu, uzticamu tematu izveide, lai tos varētu izmantot lielu akadēmisku publikāciju repozitoriju vai ziņu avotu digitālo kolekciju izpētē, ir viens no virzieniem, kas mērķtiecīgi tiek attīstīts tematu modelēšanas jomā (Chang 2009). Akadēmisko repozitoriju kontekstā neiederīgo vārdu (*intrusion words*) klātbūtne un tematu sajaukšanās (*mixing of topics*) ir nevēlama, jo mazina lietotāju paļāvību, ka modelis darbojas pareizi. Turpretī, pētot literāru darbu datus, temati nereti nav viegli interpretējami, semantiski saistītu atslēgvārdu saraksti, tomēr tie netiek noraidīti, jo var sniegt cita veida informāciju par darbu leksiku un stilistiku. Literatūras pētnieku vidū sastopams viedoklis, ka literāru tekstu izpētē ambivalenti temati var sniegt pat vairāk informācijas nekā viennozīmīgie tematu saraksti, kuriem priekšroku dotu vēsturnieki:

gadījumos, kad vārdu saraksti neveido tematiskas satura vienības, kas apraksta konkrētu referentu, tie var reprezentēt diskursu, sociolektu vai noteiktu poētiskās retorikas veidu (Underwood 2012). Citiem vārdiem sakot, tematu modelis var tikt izmantots, ne vien lai noskaidrotu, *par ko* cilvēki raksta, bet arī to – *kā* viņi raksta (Goldstone 2012).

Pieeja, kurā uz tematu saprotamību un semantisko saskanīgumu tiek likts mazāks uzsvars, pazīstama ne vien daiļliteratūras, bet arī vēstures avotu pētniecībā. Daudzos gadījumos modelēšanas rezultāti ir neviennozīmīgi un grūti interpretējami, un daži pētnieki uzsver, ka, lai arī tematu modelēšanas rezultāti ne vienmēr izmantojami kā lietošanai gatavi pierādījumi par kāda tekstu kopuma tematisko aptvērumu, tie tomēr ir lietderīgs palīglīdzeklis teksta izpētes procesā (Brett 2012). Tematu modelis var vērst uzmanību uz tēmām, kas varējušas palikt nepamanītas kvalitatīvas analīzes procesā, jo nav iespējams izlasīt visus tekstus vai izlasīt tos vienlīdz uzmanīgi (Kurvinen 2020). Tādējādi, īpaši vēsturisko un literāro avotu analīzē, tematu modeļi var papildināt hermeneitikas un tuvlasījuma pētniecības metodes, savukārt, veicot digitālu kvantitatīvu pētījumu, būtiska var būt iespēja atgriezties pie pirmavota, lai novērtētu modeļa atbilstību vai labāk izprastu, kā interpretēt rezultātus (Rhody 2012; Kurvinen 2020; Viola, Verheul 2019*)*. Minētās pieejas tematu analīzei var klasificēt kā tematu reālismu (*topic realism*) un tematu instrumentālismu (*topic instrumentalism*). Tematu reālisms ir skatījums, kurā tiek atzīts, ka modelēšanas process var tvert reprezentācijas vai teorētiskus konstruktus (rāmējumus, diskursus, naratīvus), kas reāli eksistē tekstos. Savukārt tematu instrumentālisms ir skatījums, kurā tiek atzīts, ka modeļi vienkārši sniedz informāciju par sakarībām, kas var būt noderīgas tekstu interpretācijā, neizvirzot prasību nonākt pie precīza un patiesa teksta satura realitātes atspoguļotāja modeļa (Pääkkönen, Ylikoski 2020).

Tātad ne visu veidu teksti vienlīdz labi pakļaujas mēģinājumiem formalizēt to saturu, lakoniski iekodēt jēgu dažos nozīmīgos vārdos. Tomēr neapšaubāmi tematu kvalitāte ir atkarīga arī no pārdomātas datu atlases, no teksta priekšapstrādes kvalitātes un apmācības parametriem, kas uzstādīti modelim (Wallach et al. 2009). Proti, ne katrs modelis ir uzskatāms par metodoloģiski korekti izveidotu pat tādā gadījumā, ja sniedz iedvesmojošas idejas pētniekam.

Šajā rakstā skatītā piemēra izstrādes procesā tika apsvērti vairāki scenāriji, kā LDA metodoloģija var tikt izmantota vēsturisko laikrakstu digitālās bibliotēkas izpētē. Pirmkārt, tematu modeļa apmācībai un pielāgošanai varētu tikt izmantots viss LNB digitalizētās periodikas krājums, izveidojot tematiskus blokus, kas papildinātu periodikas portāla pārlūkošanas funkcionalitāti un piedāvātu ieteikumus lietotājiem. Otrkārt, varētu tikt apmācīti un pielāgoti vairāki modeļi dažādiem LNB periodisko izdevumu segmentiem – atsevišķiem izdevumiem, izdevumu veidiem vai laika periodiem, integrējot šo informāciju kā periodikas portāla papildu funkcionalitāti. Treškārt, papildus tradicionālajiem digitālās bibliotēkas pakalpojumiem varētu tikt izveidota atsevišķa saskarne, kas ļautu lietotājam pašam atlasīt datus korpusa izveidei un veidot tematu modeli šim korpusam. Šāda pieeja sniegtu iespēju analizēt arī tikai, piemēram, kādai konkrētai tēmai vai personai veltītus materiālus. Oskara Kalpaka gadījuma izpētē lielā mērā izmantoti trešā scenārija elementi, taču piemērs sniedz noderīgas atziņas arī pārējo scenāriju īstenošanai.

Tematiska korpusa izveidei ir gan priekšrocības, gan trūkumi. No vienas puses, iepriekšēja atlase ļauj daudz detalizētāk izpētīt konkrētu interesējošo tematu. Tā, piemēram, izveidojot tematu modeli visam *Latvijas Kareivja* korpusam, tikai viens no 50 tematiem saturēja vārdu "kalpaks", savukārt Oskara Kalpaka apakškorpusā modelis ar augstāko koherences rādītāju ietver sešus tematus. No otras puses, jāpatur prātā, ka tematiska apakškorpusa izmantojums ir ierobežots, ļauj secināt tikai sākotnēji izvēlētā temata ietvaros. Izpēte, kas veikta, analizējot visu laikraksta numuru komplektu vai varbūtīgu izlasi, uzskatāma par lielākā mērā datu virzītu (*data driven*), savukārt iepriekšēja tematu atlase var palielināt pētnieka subjektivitātes (t. sk. iepriekš pieņemtu spriedumu un aizspriedumu) ietekmi modeļu veidošanā.

Oskara Kalpaka gadījuma izpētē netika ņemta vērā tematu modelēšanas laika dimensija. LDA pamata metode neņem vērā laika aspektu, t. i., laiks nav iekļauts modelī kā mainīgais, taču humanitāro zinātņu pētniecībā bieži ir nepieciešamība pētīt arhīvus un bibliotēkas, kuru materiāli publicēti vairāku gadu desmitu vai pat simtu gaitā (Marjanen 2020). Lai mazinātu šī ierobežojošā faktora ietekmi, datu kopa var tikt sadalīta secīgos segmentos, apmācot atsevišķu modeli katram segmentam. Papildinot LDA, ir tikušas izstrādātas arī citas tehnikas, kas ņem vērā laika dimensiju, piemēram, dinamisko tematu modeļu metode *(dynamic topic models*) (Blei, Lafferty 2006).

Datu kopa

Gadījuma izpētei izmantots laikraksta *Latvijas Kareivis* korpuss un no tā atvasināts Oskara Kalpaka apakškorpuss. *Latvijas Kareivis* ir oficiālais Latvijas Bruņoto spēku štāba dienas laikraksts, kas tika izdots no 1920. līdz 1940. gadam (Pētersone 1999). Līdz 1925. gadam tas tika iespiests vecajā drukā, vēlāk pāriets uz jauno ortogrāfiju. Ortogrāfijas nekonsekvence, pakāpeniskas izmaiņas un rakstības reformas ir grūtības, ar kurām bieži jāsastopas vēsturisko laikrakstu pētniekiem. Lai neradītu papildu nenoteiktību izmēģinājumu korpusā, pētījumā tika izmantota tikai *Latvijas Kareivja* modernajā ortogrāfijā iespiestā daļa. No kopējās datu kopas tika atlasīts apakškorpuss – raksti, kas satur vārdu "kalpaks".

Pulkvedis Oskars Kalpaks (1882–1919) bija Latvijas Pagaidu valdības bruņoto spēku komandieris, Landesvēra latviešu vienību un Pirmā atsevišķā latviešu bataljona komandieris (Jēkabsons 2022), kurš tiek uzskatīts par vienu no Latvijas armijas pamatlicējiem. Veicot gadījuma izpēti, hipotētiski tika pieņemts, ka ar Kalpaka vārdu varētu būt saistīti vairāki temati un tiem varētu būt mainīga aktualitāte analizētajā laika periodā. Atslēgas vārds "kalpaks" ļauj pārbaudīt arī vārdu daudznozīmības problēmas ietekmi: korpusā pieminēts Kalpaka bulvāris, Kalpaka iela, Kalpaka tilts, kas vairumā kontekstu nav tieši saistīti ar Oskara Kalpaka tematiku. Tajā pašā laikā uzvārds "Kalpaks" ir salīdzinoši rets, tādējādi nerada papildu sarežģījumus izmēģinājuma modeļa veidošanas procesā. Darba gaitā, pirmkārt, tika izveidots tematu modelis visam *Latvijas Kareivja* korpusam, otrkārt, atsevišķs tematu modelis Oskara Kalpaka apakškorpusam – pēdējais tika analizēts sīkāk.

Latvijas Kareivja korpuss satur 55,9 milj. vārdformu; Oskara Kalpaka apakškorpuss satur 1,3 milj. vārdformu. Korpusa dati ir segmentēti rakstu līmenī, tomēr rubrikas, kas satur īsas ziņas un paziņojumus, segmentēšanas procesā tikušas konsolidētas, veidojot sadaļas, kas ietver vairākas tēmas.

Teksta segmentēšanas īpatnības vai kļūdas teorētiski var kļūt par šķērsli koherenta tematu modeļa izveidei, taču, šķiet, šajā gadījumā īso ziņu savienošana nav radījusi nevēlamu ietekmi. Konsolidētās rubrikas satur daudz vienādu vārdu, saīsinājumu un skaitļu, tādējādi algoritms šos tekstus grupēja vienkopus kā atsevišķu tematu – šādi automātiski tika nošķirts materiāls, kas nesatur izvērstus pārspriedumus par tēmām, kas saistītas ar Oskara Kalpaka dzīves gaitu vai piemiņas pasākumiem. Tas vedina domāt, ka tematu modelis var būt noderīgs instruments arī pētāmā materiāla atlases un filtrēšanas procesā – lai atbrīvotos no liekajiem datiem. Vienlaikus bija novērojams, ka dažkārt atsevišķi garāki raksti savienoti kopā kļūdas dēļ: šiem tekstiem raksturīgi augsti vairāku tematu procentuālie rādītāji (tā vietā, lai izteikti dominētu viens temats).

Korpuss tika lemmatizēts, izmantojot dabīgās valodas apstrādes rīku ķēdi NLP-PIPE (Znotiņš, Cīrule 2018). Sākotnējā korpusa vārdnīca tika samazināta, atmetot vārdlietojumus, kas satur tikai vienu simbolu, – šis solis ļāva samazināt arī optiskās atpazīšanas kļūdu ietekmi. Var argumentēt, ka vārdlietojumi, kas satur tikai divus simbolus, arī var tikt atmesti, jo to vidū nav semantiski nozīmīgu vārdu, tomēr Kalpaka gadījuma izpēte parādīja, ka vismaz atsevišķos gadījumos divciparu skaitļiem bija sava loma jēgpilnas dokumentu grupēšanas procesā.

Gadījuma izpētes rezultāti

Gadījuma izpēte tikai veikta, izmantojot *Python* atvērtā koda bibliotēku *Gensim*[7]. Uzstādot modeļa apmācības parametrus, izmantoti sistēmas noklusējuma parametri un ņemti vērā *Gensim* izstrādātāju ieteikumi (Řehůřek, Sojka 2010). Tematu veidošanai izmantotais vārdu krājums (vārdnīca) tikai veidots, ņemot vērā atsevišķus vārdus, bigramas un trigramas; izmantots *bag-of-words* modelis (t. i., nav ņemta vērā vārdu secība); izveidotā vārdnīca sastāv no 5030 tekstvienībām. Sekojot *Gensim* izstrādātāju ieteikumiem, vārdnīcā tika iekļautas tekstvienības, kas lietotas vismaz 20 reižu, un netika iekļautas tekstvienības, kas sastopamas vairāk nekā 50 % dokumentu (attiecīgi vārdnīcā iekļūst maz bieži lietoto vārdu bez patstāvīgas nozīmes). Katrs LDA modelis tika apmācīts 400 iterāciju (*iterations*), 20 epohu (*epochs*) gaitā. Optimāla tematu skaita kalkulācijas balstītas (CV) koherences mērījumā.

Koherences mērījumi (*coherence measurements*) tiek izmantoti, lai novērtētu izvēlēto tematu modeļu precizitāti – piemēram, lai noteiktu, kāds tematu skaits jāizvēlas, lai

7 *Gensim: Topic Modelling for Humans*. Pieejams: https://radimrehurek.com/gensim/

veidotos koherents modelis. Šie mērījumi raksturo modeļa izvēlēto tematu saturošo vienību saskaņotību jeb vienota veseluma veidošanu starp šīm vienībām (t. i., cik lielā mērā kādā tematā ietilpstošie vārdi ir savstarpēji saistīti). Modeļu novērtēšanai tika izmantoti vairāki iepriekš zināmi un dotajā darba vidē (*Gensim* bibliotēkā) pieejami koherences mērījumi; šajā rakstā izmantoti (CV) koherences rādītāji. (CV) mērījums ir balstīts slīdošā loga (*moving window*) principā – apstrādājot kādu tematu pa segmentiem un virzoties uz priekšu pa vienam vārdam. Tiek segmentēti visbiežāk sastopamie vārdi, aprēķināta konkrētā termina atrašanas varbūtība, un izveidots mērījuma apstiprinājums; visbeidzot tiek izveidots apstrādāto segmentu (logu) rezultātu kopsavilkums. Kā parādīts pētījumos, augstākie (CV) koherences rādītāji lielā mērā sakrīt ar cilvēku subjektīvo tematu kvalitātes novērtējumu (Röder et al. 2015). Oskara Kalpaka tematu modelī augstākais (CV) rādītājs – 0,61 – tika sasniegts Oskara Kalpaka apakškorpusa modelim, kas sastāv no sešiem tematiem. Šī raksta autori subjektīvi izvērtēja četru, piecu un sešu tematu modeļu kvalitāti. 0,61 nav uzskatāms par īpaši augstu modeļa koherences rādītāju, un arī subjektīvā vērtējumā nevarēja apgalvot, ka visi apakškorpusa raksti pārliecinoši tika sagrupēti pa tematiem.

Piemēra izstrāde bija balstīta hipotētiskos pētnieciskos jautājumos: kādi temati ir saistīti ar Oskara Kalpaka vārdu laikrakstā *Latvijas Kareivis* laika posmā no 1925. līdz 1940. gadam? Kādos kontekstos tiek minēts Kalpaka vārds? Cik daudz dažādu tematu ir saistīti ar Kalpaku? Kā šo tematu popularitāte mainās laika gaitā?

Pārlūkojot *Latvijas Kareivja* korpusa 50 tematu modeli (tematu izlasi skat. 1. tabulā), varam secināt, ka tajā spilgti iezīmējas temati, kas saistīti, piemēram, ar sportu, izglītību, transporta pārvadājumiem. Lielā skaitā atrodami temati, kas satur valstu un tautību nosaukumus, turklāt nereti vairākas valstis tiek grupētas viena temata ietvaros. Šķiet, šāds tematisks dalījums varētu būt noderīgs, lai virzītu lasītāju pie viņam aktuāliem rakstiem, tomēr būtu nepieciešams padziļināts pētījums, lai secinātu, piemēram, cik lielā mērā tematos, kuros minētas vairākas valstis, vērojama tematu sajaukšanās, kā arī vai sastopami neiederīgie vārdi. Rūpīgāka modeļa kvalitatīva izpēte varētu sniegt atbildi uz jautājumu, vai šie tematiskie ietvari (tematu atslēgvārdu saraksti paši par sevi) var kalpot par informācijas avotu, pētot, kādā kontekstā tiek runāts par dažādām valstīm: vai būtu pamatoti sacīt, ka Igaunija un Somija galvenokārt tiek saistītas ar ciešām diplomātiskām attiecībām un vizītēm, savukārt Holande un Anglija – ar finanšu jautājumiem utt.

Vārds "Kalpaks" 50 tematu modelī parādās vienā no tematiem. Jāņem vērā, ka pēc nozīmīguma Kalpaks šeit nav viens no galvenajiem temata vārdiem – tā, piemēram, ja lietotājam, strādājot ar tekstiem digitālā kolekcijā, būtu redzami tikai pieci vai septiņi temata nozīmīgākie vārdi, Kalpaka to vidū nebūtu vispār. Pastāvošajā rāmējumā varam spriest drīzāk par to, ka tādi jēdzieni kā *Rīga, latvietis, vēsture, ordenis* un *novembris* saistās arī ar Kalpaka vārdu, nevis pretēji. Kāpinot tematu skaitu, Kalpaka vārds varētu parādīties vairākos tematos, taču, izvēloties lielāku tematu skaitu šim korpusam, samazinās modeļa (CV) koherences rādītājs.

2. temats	armija, aizsardzība, vienība, kars, apmācība, zirgs, militārs, valsts_aizsardzība, laiks, organizācija, darbība, uzdevums, dienests, karaspēks, ierocis, pulkst_19, grupa, jātnieks, tikt, vadība, kā, sastāvs, daļa, sagatavošana, kauja, manevrs, viss, rezerve, sakars, kurš
3. temats	savienība, padome, tauta, kongress, Rumānija, tauta_savienība, Turcija, turks, Austrija, ungārs, Dienvidslāvija, antante, Danciga, Bulgārija, starptautisks, Grieķija, loceklis, tēvzeme_mīlestība, Vīne, Kronvalds, mazs, delegāts, Bukaresta, grieķis, bulgārs, pārstāvis, ukrainis, sanākt, Eiropa, ārlietas
4. temats	Latvija, Igaunija, Somija, ārlietas, Baltija, mēs, igaunis, vakar, sūtnis, valsts, pārstāvis, ministrija, ārlietas_ministrs, Tallina, Rīga, Baltija_valsts, ierasties, prese, izbraukt, direktors, vadītājs, delegācija, iepazīties, ārlietas_ministrija, piedalīties, konference, sūtniecība, apmeklēt, iepazīties_ar, notikt
6. temats	tanks, varēt, mašīna, ātrums, ceļš, mm, izmantot, vai, automobilis, degviela, čemberlens, gāze, smags, šāds, līdzeklis, ierocis, viegls, lietot, auto, svars, iespēt, šis, veids, transports, nafta, aparāts, cm, kustība, katrs, dienvidaustrumi
9. temats	zviedrs, Holande, Beļģija, Zviedrija, soms, Šveice, beļģis, soma, dānis, Vācija, marka, Francija, vicepriekšsēdētājs, lēdija, angļu_franču, birža, vāci, Anglija, Itālija, anglis_vēstnieks, franks, Rīga_birža, ražība, darbs_ražība, tīrs_peļņa, kurss, Amsterdama, padome_sesija, viceadmirālis, beigas
12. temats	Rīga, satiksme, vilciens, stacija, līnija, pasts, pasažieris, vagons, akc, akc_sab, autobuss, 13_05, sab, ceļš, līdz, virsvalde, biļete, pa, telegrāfs, pienākt, prece, brauciens, braukt, starp, tarifs, jūrmala, Jelgava, aiziet, Lielupe, pārvadāt
16. temats	sports, vienība, sacīkste, sacensība, spēle, sek, Latvija, uzvarēt, notikt, pirmais, labs, cīņa, mēs, min, vieta, uzvara, Rīga, svars, laukums, punkts, bet, futbols, savienība, balva, vārti, US, gūt, valsts, LSB, sportists
17. temats	skola, pamatskola, izglītība, skolotājs, kurss, mācība, jaunatne, skolēns, ģimnāzija, audzēknis, direktors, valsts, beigt, izglītība_ministrs, institūts, ministrija, izglītība_ministrija, lauksaimniecība, vidusskola, akadēmija, klase, jauns, audzināšana, pārbaudījums, skola_jaunatne, skolnieks, bērns, darbs, arodskola, praktisks
32. temats	gads, dzimt, Latvija, Rīga, latvietis, vēsture, ordenis, novembris, mirt, viņš, kars, pirmais, cīņa, 1919, gaidīt_laiks, atbrīvošana, janvāris, maijs, pēc, kā, krievs, līdz, pie, marts, laiks, oktobris, strēlnieks, februāris, Kalpaks

1. tabula. Tematu izlase no *Latvijas Kareivja* korpusa 50 tematu modeļa; norādītas katra temata 30 raksturīgākās tekstvienības.

Oskara Kalpaka apakškorpusa sešu tematu modelī 1., 3. un 4. temata vārdu saraksti apstiprināja gaidas, ka temati būs saistīti ar militāriem terminiem, proti, tiek pieminēta armija, pulks, cīņa, ģenerālis, pulkvedis u. c. (skat. 2. tabulu). Savukārt 2., 5. un 6. temats rāda jēdzienus, kas mazākā mērā saistās ar kara lietām: *iela, pilsēta, pulkst, skola* u. c. Novērtējot subjektīvi, katrs no tematiem šķiet veidojam kādu konkrētu tematisku identitāti, tomēr īpaši 2., 5. un 6. temata gadījumā nav skaidrības, kā šie jēdzieni saistīti ar pulkvedi Kalpaku, un ir nepieciešams vērsties pie rakstu pilnajiem tekstiem, lai saprastu, kā šos tematus interpretēt.

1. temats	tauta, arī, šī, savs, cīņa, varēt, armija, vēl, tad, kad, jūs, tikai, zeme, jo, jau, karavīrs, spēks, daudz, latvietis, visa, Iatvis, vai, es, labs, pats, Aris[8], viņa, ja, dzīve, viens
2. temats	iela, pilsēta, Aris, valde, vakar, vieta, pa, ministrija, biedrība, 10, paredzēt, ls, ministrs, jau, nodaļa, notikt, daļa, galva, pīkstēt[9], kāds, vēl, telpa, varēt, nolemt, vai, nams, 000, atrast, vakars, policija
3. temats	ministrs, ģenerālis, svētki, pilsēta, aizsargs, prezidents, pulks, armija, karavīrs, priekšnieks, piemiņa, kaps, komandieris, piemineklis, pulkv, pulkvedis, arī, notikt, valsts_prezidents, krist, garnizons, vieta, baznīca, biedrība, svinība, Liepāja, svinīgs, bataljons, organizācija, dievkalpojums
4. temats	rota, pulks, bataljons, kauja, armija, pulkvedis, marts, 1919, lielinieks, komandieris, janvāris, uzbrukums, pulka, Cēsis, jātnieks, virsnieks, karavīrs, vāci, tikt, daļa, kājnieks, karaspēks, ienaidnieks, ieņemt, vienība, atsevišķs, cīņa, eskadrons, jau, muiža
5. temats	pulkst, 30, pīkstēt, 20, koncerts, 19, ziņa, 15, 18, 12, 10, 17, 00, šodien, skaņa, 22, 16, plate, rīts, vakars, 13, iela, mūzika, dziesma, pl, pilsēta, 21, piedalīties, opera, koris
6. temats	ls, pag, 10, skola, pamatskola, 10_ls, 50, kl, 25, 20, 100, 50_ls, pagasts, ba, valde, grāmata, darbinieks, skolotājs, 000, mazpulks, skolēns, 20_ls, 15, 100_ls, sab, 25_ls, 30, pils, pilsēta, biedrība

2. tabula. Oskara Kalpaka apakškorpusa sešu tematu modelis; norādītas katra temata 30 raksturīgākās tekstvienības.

Subjektīvi pārskatot atbilstošos rakstus[10], ir novērojams, ka 1., 3. un 4. temats dominē rakstos, kas veltīti pulkvedim Oskaram Kalpakam un ar viņa personību saistītām norisēm. Šie temati nošķir vairākus atšķirīgus kontekstus: Oskara Kalpaka atceres dienām tapuši raksti, kuros paustas pārdomas par Kalpaka nozīmi Latvijas vēsturē (1. temats); Oskara Kalpaka atceres dienu notikumu pārskati, kuros aprakstītas dažādas svinības un ceremonijas (3. temats); raksti, kuros tiek pārspriesti 1919. gada notikumi – Latvijas Bruņoto spēku izveide un kaujas pret lieliniekiem (4. temats).

5. temats dominē paziņojumos un reklāmās par koncertiem un citiem kultūras pasākumiem. Atsevišķos gadījumos šeit var būt pieminēti Kalpakam veltīti pasākumi, taču vairumā gadījumu tiek minēts Kalpaka bulvāris un Kalpaka iela kā norises vietas. Šajā tematā sastopamie

8 Vārds "Aris" tematu vārdnīcā ieviesies sistemātiskās optiskās atpazīšanas kļūdas dēļ. Vārds "arī" nereti atpazīts kā "ari", turklāt "Ari" bieži sastopams teikuma sākumā, rakstīts ar lielo sākumburtu. Morfoloģiskās marķēšanas rīka interpretācijā tas kļuvis par īpašvārdu – "Aris".

9 Vārds "pīkstēt" tematu vārdnīcā ieviesies sistemātiskas optiskās atpazīšanas kļūdas dēļ. Saīsinājums "plkst." neprecīzi atpazīts kā "pīkst", savukārt morfoloģiskās marķēšanas rīks to pārveidojis pamatformā – "pīkstēt".

10 Oskara Kalpaka apakškorpusa rakstu pilnie teksti pieejami šeit: https://doi.org/10.5281/zenodo.6569249. Raksti sagrupēti pa tematiem, ņemot vērā tematu ar visaugstāko rādītāju.

skaitļi ir norises datumi un laiki; vairākās variācijās sastopams vārds "pulksten" (skat. arī 9. beigu piezīmi). 6. temats dominē galvenokārt paziņojumos, kas saistīti ar skolām, kuras nosauktas Oskara Kalpaka vārdā. Parasti šīs ziņas saistītas ar ziedojumu vākšanu un dāvinājumiem skolām. Kopš 1939. gada šis temats sastopams arī paziņojumos par ziedojumiem Latvijas aizsardzības spēkiem – skaitļi ir ziedotās summas. 2. temats satur vislielāko atšķirīgas tematikas atslēgvārdu sajaukumu. Šis temats dominē rakstos, kas dažkārt atsaucas uz Oskaru Kalpaku, bet lielā daļā gadījumu tiek pieminēts tvaikonis "Kalpaks". Vienā no rakstiem minēta persona ar uzvārdu "Kalpaks", kura nav Oskars Kalpaks.

1. attēlā redzamā vizualizācija apstiprina novērojumus, kas gūti tekstu un tematu subjektīvā analīzē – raksti, kuros dominē 1., 3. un 4. temats, saturiski ir vairāk saistīti, un tie kā līdzīgi tuvāk sagrupēti arī daudzdimensiju kartējumā, savukārt 5. un 6. temats saturiski nav saistīti ar Oskaru Kalpaku, un arī kartējumā tie atrodas perifērijā.

LDA modelī katrs raksts var ietvert vairākus tematus dažādās nozīmīguma pakāpēs. Raksti, kuru tematiskajā kompozīcijā sajaukti vairāki temati ar lielu nozīmīguma pakāpi, biežāk izrādās vai nu Kalpaka tēmai nepiederīgi, vai tādi, kuros garāki raksti savienoti kopā segmentēšanas kļūdas dēļ. Daudzos gadījumos tomēr viena temata īpatsvars ir izteikti dominējošs – virs 80 %. Tā, piemēram, Kalpaka piemiņai veltīta publikācija ar 83 % 1. temata īpatsvaru vēsta:

> *Svinot ik gadus Kalpaka bataljona gada svētkus mēs pieminam to laikmetu, kurš mūsu armijas un līdz ar to mūsu tautas vēsture ierakstīts neizdzēšamiem nacionālo varoņu asinīm slacītiem burtiem. Katras tautas vēsture ir šāds laikmets, atmiņa par kuru iet no paaudzes uz paaudzi, paužot par to spēku, kurš atsvabināja tautu no svešas varas važām*[11].

Raksti ar lielu 1., 3. vai 4. temata īpatsvaru drīzāk izvērsti vēsta par jautājumiem, kas saistīti ar Oskaru Kalpaku vai viņam veltītiem pasākumiem, tomēr tā nav vienmēr. Tā, piemēram, kāda publikācija ar 85 % 1. temata īpatsvaru veltīta Zigfrīda Meierovica (1887–1925) atcerei, un Kalpaks tajā ir tikai pieminēts (tajā pašā laikā tematiski šis raksts tuvs patriotiskajiem Kalpakam veltītajiem rakstiem):

Īstā brīdī laime mums sūtīja Kalpaku kara frontē un Meierovicu vēl grūtāka ārējās politikas frontē. Tos abus ņēma tad, kad viņu darbs vēl ārkārtīgi vajadzīgs. Mums atliek tikai ticēt Latvijas zvaigznēm kā to darīja šie divi Latvijas krietnākie dēli[12].

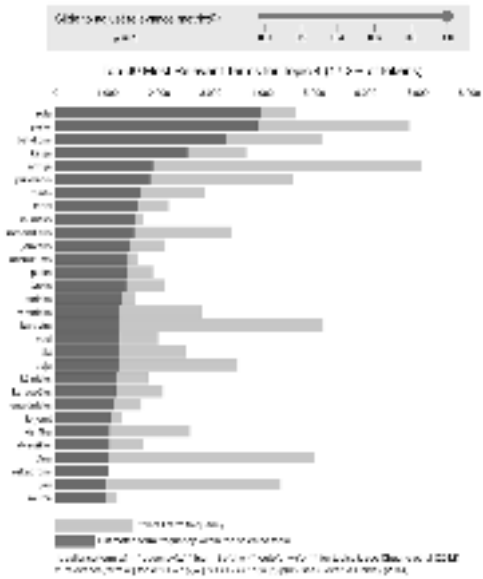
Šie novērojumi vedina domāt, ka, veidojot kādai tēmai veltītu korpusu, varētu būt lietderīgi atlasīt rakstus, kuros interesējošais vārds minēts vairāk nekā vienu reizi – lai atsijātu maznozīmīgākus pieminējumus.

11 E.G. Kalpaku pieminot. *Latvijas Kareivis*, 03.03.1925., 1. lpp. Rakstus, kuros dominē pirmais temats, datu kopā sk. pie "Topic 0": https://doi.org/10.5281/zenodo.6569249

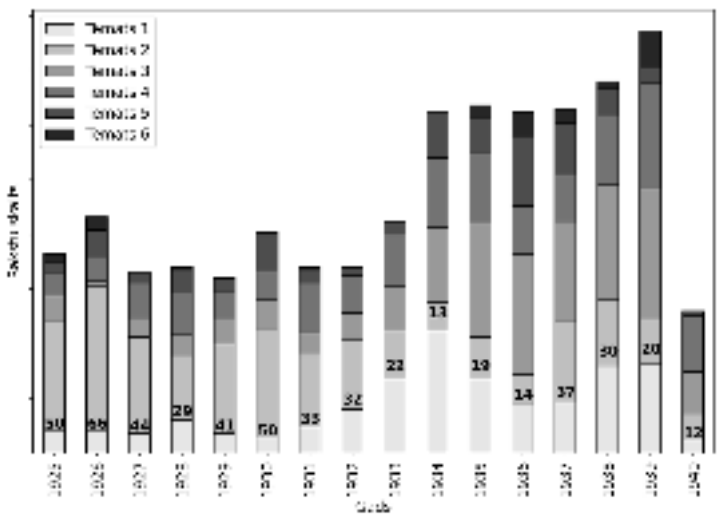
12 Gailītis, M. Valsts vīri par pirmā diplomāta nāvi. *Latvijas kareivis*, 26.08.1925., 3. lpp.

1. attēls. Kreisajā pusē: sešu Kalpaka apakškorpusa tematu daudzdimensiju kartējums. Labajā pusē: 30 nozīmīgākie ceturtā temata jēdzieni.

Vizualizācija veidota, izmantojot *pyLDAvis Python* bibliotēku (Sievert, Shirley 2014).



2. attēls. Sešu tematu sadalījums Oskara Kalpaka apakškorpusa rakstos no 1925. līdz 1940. g.



2. attēlā aplūkojams katram tematam veltīto rakstu skaits laikposmā no 1925. līdz 1940. gadam[13]. Kopumā Kalpaka pieminējumu skaits izteikti palielinās, sākot ar 1934. gadu, īpaši liels pieaugums vērojams 1. un 3. tematam, savukārt 2. tematam vērojams samazinājums; 6. temats īpaši aktuāls 1939. gadā. Šīs izmaiņas, iespējams, ir interpretējamas 1934. gada valsts apvērsuma kontekstā – lai to noskaidrotu, būtu jāveic kvalitatīva rakstu izpēte. Jāpatur prātā, ka gadījuma izpētēs netika nošķirti un saskaitīti īstie un neīstie Kalpaka pieminējumi, tādējādi, lai nonāktu pie precīzākiem datiem par Kalpaka tematu popularitāti, būtu jāturpina pieminējumu analīzes un filtrēšanas darbs.

Secinājumi

Gadījuma izpēte liecina, ka LDA tematu modelēšanas metodoloģija ir noderīga vēsturiskās periodikas pētniecībai un varētu būt piemērota, lai to integrētu kā jaunu funkcionalitāti vai papildu saskarni LNB digitālajās kolekcijās. LDA metodoloģija ir daudzkārt pārbaudīta lietojumiem citās valodās, un tā tiek uzskatīta par īpaši piemērotu akadēmisku publikāciju, žurnālu un laikrakstu tematu modelēšanai. Papildus šajā rakstā skatītajam LDA variantam pasaulē tikuši izstrādāti risinājumi, kas ļauj precīzāk tvert tematu izmaiņas ilgstošā laika periodā.

Gadījuma izpētes gaitā tika definēta darbplūsma, kas nepieciešama, lai izveidotu uzticamu tematu modeli: datu priekšapstrāde (korpusa kompilēšana, tīrīšana, morfoloģiskā marķēšana), modeļa iteratīva apmācība, provizoriski izveidojot modeļus ar dažādu tematu skaitu, tematu modeļa koherences mērīšana, tematu sadalījuma subjektīva izvērtēšana, optimālā modeļa izvēle, vizualizāciju un citu modeļa reprezentāciju veidošana. No lietotāja viedokļa, izstrādājot atbilstoša pētniecības pakalpojuma risinājumu, vēlams nodrošināt, lai lietotājs var piekļūt modeļa avottekstiem.

Oskara Kalpaka apakškorpusa tematu modelis apliecināja, ka LDA ļauj veidot semantiski saskanīgus, noderīgus tematus, tomēr, lai pilnībā interpretētu rezultātus, nepieciešams vērsties pie pašiem rakstiem. Rakstu turpmāka subjektīva pārbaude atklāja, ka tematu dalījums ļāvis jēgpilni nošķirt rakstus, tomēr sastopama arī neatbilstoša attiecināšana. Secināms, ka Oskara Kalpaka gadījuma izpētē lietotā metode veiksmīgi izmantojama tematu instrumentālisma pieejā, savienojot gan kvalitatīvās, gan kvantitatīvās metodes avotu izpētē. Turpmāka modeļa uzstādījumu pielāgošana un izpēte būtu nepieciešama, lai LDA metodoloģiju varētu izmantot tematu reālisma pieejā vai situācijās, kurās pētniekiem ir ierobežotas iespējas pārlūkot rakstus un mazāk informācijas par to, kā modelis veidots un kādi ir tā ierobežojumi.

Izvēloties modeļa apmācības uzstādījumus, tikai atmesti vārdi, kas lietoti vairāk nekā 50 % tekstu, tādējādi tikai izslēgta lielākā daļa palīgvārdu, vietniekvārdu, bieži lietotu apstākļa vārdu. Lietojumu scenārijos, kuros digitālas kolekcijas lietotājam būtisks tieši raksta priekšmets

13 Raksti tika skaitīti, par pamatu ņemot to tematu, kuram rakstā ir augstākā procentu vērtība.

(referents) un redzami nevis 30, bet tikai daži temata atslēgvārdi, iespējams, būtu ieteicams pielāgot uzstādījumus tādējādi, lai modelī būtu vēl mazāk vārdu bez patstāvīgas nozīmes vai tiktu iekļauti tikai lietvārdi. Savukārt lietojumu scenārijos, kuros pētnieciskie mērķi ir saistīti arī ar tekstu stilistikas un retorikas pētniecību, vārdšķiru daudzveidība paturama. Kā parādīja Oskara Kalpaka modeļa piemērs, skaitļu iekļaušana modelī ne vien nodrošina, ka saglabājas tematam nozīmīgi gadskaitļi (kā 1919. gads), bet ļauj arī identificēt materiālus, kurus pētnieks, iespējams, vēlētos izņemt no pētāmo datu kopas.

Turpinot metodes izstrādi, būtu nepieciešams turpināt pielāgot LDA tematu modeļa uzstādījumus, analizējot arī citas datu kopas, kā arī ieteicams LDA rezultātus salīdzināt ar citu algoritmu veiktspēju.

Bibliogrāfija

Abney, Steven, Bird, Steven (2010). The Human Language Project: building a universal corpus of the world's languages. *Proceedings of the 48th Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, pp. 88–97.

Alabi, Jesujoba, Amponsah-Kaakyire, Kwabena, Adelani, David, et al. (2020). *Massive vs. Curated Embeddings for Low-Resourced Languages: the Case of Yorub`a and Twi. Proceedings of the 12th Language Resources and Evaluation Conference*. European Language Resources Association, pp. 2754–2762.

Alves, Diego, Thakkar, Gaurish, Tadić, Marko (2020). Evaluating Language Tools for Fifteen EU-official Under-resourced Languages. *Proceedings of the 12th Language Resources and Evaluation Conference.* European Language Resources Association, pp. 1866–1873.

Baklāne, Anda, Saulespurēns, Valdis (2022). The application of latent Dirichlet allocation for the analysis of Latvian historical newspapers: Oskars Kalpaks' case study. *Nauka. tehnologii, innovacii,* No. 1(21), s. 29–37.

Blei David M., Lafferty, John D. (2007). A correlated topic model of Science. *Annals of Applied Statistics*, Vol. 1(1), pp. 17–35.

Blei, David M., Lafferty, John D. (2006). Dynamic topic models. *Proceedings of the 23rd international conference on Machine Learning*, pp. 113–120.

Blei, David M., Ng, Andrew Y., Jordan, Michael I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3 (January), pp. 993–1022.

Blei, David (2012). Topic modeling and digital humanities. *Journal of Digital Humanities*, Vol. 2, No. 1, pp. 8–12. Available: http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/ [accessed 18.06.2022.].

Block, Sharon (2006). Doing More with Digitization: An introduction to topic modeling of early American sources. *Common-place: The Interactive Journal of Early American Life*, 6.2. Available: http://commonplace.online/article/doing-more-with-digitization/ [accessed 18.06.2022.].

Bollmann, Marcel (2019). A Large-Scale Comparison of Historical Text Normalization Systems. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 1. Association for Computational Linguistics, pp. 3885–3898.

Brett, Megan R. (2012). Topic Modeling: A Basic Introduction. *Journal of Digital Humanities*, Vol. 2, No. 1, pp. 1–2. Available: http://journalofdigitalhumanities.org/2-1/topic-modeling-a-basic-introduction-by-megan-r-brett/ [accessed 18.06.2022.].

Chang, Jonathan, Boyd-Graber, Jordan, Gerrish, Sean, et al. (2009). Reading Tea Leaves: How Humans Interpret Topic Models. *Advances in Neural Information Processing Systems* 22. Available: https://proceedings.neurips.cc/paper/2009/file/f92586a25bb3145facd64ab20fd554ff-Paper.pdf [accessed 18.06.2022.].

Ehrmann, Maud, Romanello, Matteo, Clematide, Simon, et al. (2020). Language Resources for Historical Newspapers: The Impresso Collection. *LREC 2020 Proceedings*, pp. 958–968.

Goldstone, Andrew, Underwood, Ted. (2012). What Can Topic Models of PMLA Teach Us About the History of Literary Scholarship? *Journal of Digital Humanities*, Vol. 2, No. 1, pp. 39–48. Available: http://journalofdigitalhumanities.org/2-1/what-can-topic-models-of-pmla-teach-us-by-ted-underwood-and-andrew-goldstone/ [accessed 18.06.2022.].

Hall, David, Jurafsky, Daniel, Manning, Christopher D. (2008). Studying the history of ideas using topic models. *Proceedings of the 2008 conference on empirical methods in natural language processing*, pp. 363–371.

Hengchen, Simon (2017). *When Does it Mean? Detecting Semantic Change in Historical Texts*. Ph.D. thesis. Université libre de Bruxelles.

Jēkabsons, Ēriks (2022). Oskars Kalpaks. *Nacionālā enciklopēdija*. Pieejams: https://enciklopedija.lv/skirklis/26024-Oskars-Kalpaks [skatīts 18.06.2022.].

Krūmiņa, Līga (2012). Digitalizācija Latvijā pasaules pieredzes kontekstā. *Bibliotēku pasaule*, Vol. 57, 39.–45. lpp.

Kurvinen, Heidi (2020). Towards Digital Histories of Women's Suffrage Movements. Fridlund, Matts, Oiva, Mila, Paju, Petri (eds.) *Digital Histories: Emergent Approaches within the New Digital History*. Helsinki University Press, pp. 149–163.

Marjanen, Jani, Zosa, Elaine, Hengchen, Simon, et al. (2020). Topic

Modelling Discourse Dynamics in Historical Newspapers. *Post-Proceedings of the 5th Conference Digital Humanities in the Nordic Countries (DHN 2020)*, pp. 63–77.

McGillivray, Barbara (2021). Computational methods for semantic analysis of historical texts. Kristen Schuster, Stuart Dunn. *Routledge International Handbook of Research Methods in Digital Humanities.* London; New York: Routledge, Taylor & Francis Group, pp. 261–274.

Nelson, Robert K. (2011). *Mining the Dispatch*. Available: https://dsl.richmond.edu/dispatch/introduction [accessed 18.06.2022.].

Newman, David, Chemudugunta, Chaitanya, Smyth Padhraic, et al. (2006). Analyzing entities and topics in news articles using statistical topic models. *Intelligence and Security Informatics, IEEE International Conference on Intelligence and Security Informatics*, pp. 93–103.

Pääkkönen, Juho, Ylikoski, Petri (2020). Humanistic interpretation and machine learning. *Synthese*, 199 (Sept.), pp. 1461–1497.

Pētersone, Inta (1999) (red.). Latvijas Kareivis. *Latvijas Brīvības cīņas 1918–1920: enciklopēdija*. Preses Nams, 187. lpp.

Řehůřek, Radim, Sojka, Petr (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Available: http://is.muni.cz/publication/884893/en [accessed 18.06.2022.].

Rhody Lisa M. (2012). Topic Modeling and Figurative Language. *Journal of Digital Humanities*, Vol. 2, No. 1. Available: http://journalofdigitalhumanities.org/2-1/topic-model-data-for-topic-modeling-and-figurative-language-by-lisa-m-rhody/ [accessed 18.06.2022.].

Röder, Michael, Both, Andreas, Hinneburg, Alexander (2015). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining WSDM '15*, pp. 399–340.

Sievert, Carson, Shirley, Kenneth (2014). LDAvis: A method for visualizing and interpreting topics. *Proceedings of the workshop on interactive language learning, visualization, and interfaces*. Association for Computational Linguistics, pp. 63–70.

Skadiņa, Inguna, Veisbergs, Andrejs, Vasiļjevs, Andrejs et al. (2012). The Latvian Language in the Digital Age / Latviešu valoda digitālajā laikmetā. *META-NET White Paper Series: Latvian*. Berlin: Springer.

Templeton, Thomas C., Brown, Travis, Battacharyya, Sayan, et al. (2011). Mining the Dispatch under Supervision: Using Casualty Counts to Guide Topics from the Richmond Daily Dispatch Corpus. *Chicago Colloquium on Digital Humanities and Computer Science*.

Ūdre, Dace, Baltiņa, Dagnija et al. (2019). *Digital Approaches in Cultural Heritage: towards a pan-Baltic cooperation network: final report*. Riga: National Library of Latvia. Available: https://dom.lndb.lv/data/obj/781145.html [accessed 15.09.2022.].

Underwood, Ted (2012). *Topic modeling just made simple enough*. Blog post. Available: https://tedunderwood.com/2012/04/07/topic-modeling-made-just-simple-enough/ [accessed 18.06.2022.].

Vīksna, Rinalds, Kirikova, Marite, and Kiopa, Daiga (2020). Exploring the Use of Topic Analysis in Latvian Legal Documents. *COUrT - CAiSE for Legal Documents*, *Virtual Workshop*. Available: http://ceur-ws.org/Vol-2690/COUrT-paper4.pdf [accessed 18.06.2022.].

Viola, Lorella, Verheul, Jaap (2019). Mining ethnicity: Discourse-driven topic modelling of immigrant discourses in the USA, 1898–1920. *Digital Scholarship in the Humanities*, Vol. 35(4), pp. 921–943.

Wallach, Hanna, Mimno, David, McCallum, Andrew (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems*, Vol. 23 (January), pp. 1973–1981.

Znotiņš, Artūrs, Cīrule, Elita (2018). NLP-PIPE: Latvian NLP Tool Pipeline. Human Language Technologies. *The Baltic Perspective*, IOS Press, Vol. 307, pp. 183–189.

Zariņš, Uldis (2014). Eiropas kultūras mantojums digitālajā vidē. *Latvijas intereses Eiropas Savienībā,* No. 2, 41.–55. lpp. Pieejams: https://dom.lndb.lv/data/obj/61436.html [skatīts 15.09.2022.].

Summary

The Model of Latent Dirichlet Allocation in the Topic Analysis of Latvian Soldier: Oskars Kalpaks' Case Study

Anda Baklāne, Valdis Saulespurēns

Keywords: topic modelling, digitized newspapers,
digital history, topic coherence, National Library of Latvia

The paper presents a case study of the application of the LDA (latent Dirichlet allocation) model for the analysis of topics in the corpus of the historical daily newspaper of Latvian armed forces *Latvian Soldier* (1925–1940). Although topic modelling is one of the most popular techniques for analysing text in digital humanities, this methodology has not been extensively tested for texts in Latvian. The case study was conducted to explore the possibilities for implementing topic models as new functionality for exploring newspapers in the digital library of the National Library of Latvia. To imitate different use cases of topic modelling, two models were created: a model consisting of 50 topics for the whole corpus of the *Latvian Soldier*, as well as a six-topic model of the subcorpus compiled from articles that contain the name 'Kalpaks'. It was demonstrated that both models produced usable, semantically coherent topics that could aid the exploration of historical newspapers. It was concluded that the quality of the models in the current state was sufficient to follow the approach of topic instrumentalism, which views topics as incomplete representations of texts that are a useful augmentation of the investigative process. The acquired topic models seem particularly useful for combining research practices of distant and close reading. Further testing and adjustment of the parameters are needed to produce concise and unambiguous topics that could be reliably used in research situations where extensive analysis of the sources and verification is not expected.