Ülevaade Põhja- ja Baltimaade digihumanitaaria konverentsist Tartus

HANNA-RIIN KARU, GERTH JAANIMÄE, AIGI RAHI-TAMM, TOIVO KIKKAS, SVEN LEPA

2025. aasta märtsi alguses toimus Tartus Eesti Rahva Muuseumis Põhja- ja Baltimaade digihumanitaaria võrgustiku (DHNB) aastakonverents, mis oli järjekorras juba üheksas ja kandis sedapuhku pealkirja „Digitaalsed unistused ja praktikad" („Digital Dreams and Practices").[1] Konverentsi ettekanded, mida oli kokku üle saja, keskendusid tehisintellekti (TI) kiiresti arenevale potentsiaalile, selle lõimimisele traditsiooniliste humanitaarteadustega ja rollile akadeemiliste piiride ületamisel.

Sisukad ettekanded pälvisid tähelepanu nii digihumanitaaria valdkonnaga alles tutvust tegevale osalejale kui ka kogenud spetsialistidele. Laias laastus võib digihumanitaariaga seotud teemad jaotada tööprotsesside lõikes järgmiselt: ainese digiteerimine ja digitaalselt kättesaadavaks tegemine, teabe otsitavaks muutmine, selle analüüs ja visualiseerimine ning uute tööriistade loomine. Kuna iga etapp nõuab spetsiifilisi oskusi, aega ja ressursse, keskenduvad digihumanitaariaga seotud uurimisprojektid sageli vaid ühele probleemile ega võta eesmärgiks tegeleda neist kõigiga korraga. Tulemuste usaldusväärsus ja erinevate meetodite kitsaskohad olidki konverentsil läbivaks aruteluteemaks. Neid käsitles oma avaettekandes ka Tartu Ülikooli digihumanitaaria külalisprofessor Maciej Eder, kes juhtis tähelepanu tõsiasjale, et ka piiratud andmevalimi uurimisel võib saada väärtuslikke ja laiema tähtsusega tulemusi.

Allikatest teabe otsitavaks muutmine

Mitmed ettekanded keskendusid ajalooallikate sisu masinloetavaks muutmisele. Kuigi arhiivid on tegelenud intensiivselt arhiiviainese digiteerimisega, mis lihtsustab oluliselt neile juurdepääsu, ei saa üldjuhul veel otsingut teostada allikate sisus.

Hanna-Riin Karu, magistrant, projekti spetsialist, ajaloo ja arheoloogia instituut, Tartu Ülikool, Ülikooli 18, Tartu 50090, hanna-riin.karu@ut.ee

Gerth Jaanimäe, doktorant, eesti ja üldkeeleteaduse instituut, Tartu Ülikool, Ülikooli 18, Tartu 50090, gerth.jaanimae@gmail.com

Aigi Rahi-Tamm, PhD, arhiivinduse osakonna juhataja, arhiivinduse professor, ajaloo ja arheoloogia instituut, Tartu Ülikool, Ülikooli 18, Tartu 50090, aigi.rahi-tamm@ut.ee

Toivo Kikkas, PhD, arhiivinduse teadur, ajaloo ja arheoloogia instituut, Tartu Ülikool, Ülikooli 18, Tartu 50090, toivo.kikkas@ut.ee

Sven Lepa, MA, Tartu kasutusosakonna asejuhataja, Rahvusarhiiv, Nooruse 3, Tartu 50411, sven.lepa@ra.ee

[1] https://dhnb.eu/conferences/dhnb2025/

Viimastel aastatel on TI toel kiirelt arendatud lahendusi ajalooliste käsikirjaliste tekstide tuvastamiseks. Tuntuim neist on Eestiski rakendamist leidnud platvorm Transkribus,[2] mille kasutusvõimalustest kõneles Hollandi Huygensi instituudi teadur ja Transkribust arendava kooperatiivi READ-COOP kogukonnajuht C. Annemieke Romein.

Suuri edusamme on käsikirjalise teksti tuvastuse võimekuse arendamisel teinud Rootsi Riigiarhiiv. Erik Lenas ja Olof Karsvall tutvustasid tekstituvastuseks treenitud mudelit Swedish Lion, mis suudab aastatest 1600–1900 rootsikeelseid allikaid tuvastada umbes 5% CER-i ehk tähemärgi veamääraga, mis on vägagi hea tulemus. Selleni jõudmiseks tuli mudeli õpetamiseks luua kõigepealt käsitsi treeningmaterjal – mõnikümmend andunud arhiivikasutajat kirjutas ümber 1,5 miljonit tekstirida –, mille põhjal töötati tuvastusprotsessi jaoks välja terviklik töövoog HTRFlow,[3] mis on paindlik ja avatud koodiga. Selle abil on võimalik välja lugeda nii vabas kui ka struktureeritumas vormis koostatud tekste. Praeguseks on tuvastatud 7,2 miljonit kaadrit jooksvalt kirjutatud teksti ja 20 miljonit kaadrit struktureeritud sisuga allikaid.[4]

Andmete tõlgendamine

Eraldi tähelepanu pöörati TI rakendustele ning käsitleti andmete kättesaadavuse, tõlgendamise ja eetika küsimusi. Samuti tõsteti esile digitaalse kultuuripärandi ühiskondlikku potentsiaali ning vajadust vastutustundlike ja kaasavate uurimispraktikate järele.

Tartu Ülikooli tehisintellekti professor ja Eesti TI tippkeskuse juht Meelis Kull rääkis suurte keelemudelite taustal olevatest protsessidest ja rõhutas, kui oluline on TI süsteemi otsuste läbipaistvuse ja seletatavuse teadvustamine. Ta tuletas meelde ka üht olulisimat põhimõtet: TI-d kasutades on kõige alustalaks õigesti sõnastatud tööülesanne (õigesti esitatud küsimus) ehk viip.

Tekstituvastuse vallas on välja arendamisel suured multimodaalsed keelemudelid (MLLM) ehk generatiivsed tehisintellekti mudelid, mis suudavad korraga analüüsida näiteks teksti ja pilti. Yunting Xie Uppsala ülikoolist esitles aastatest 1945–1975 pärinevate Rootsi patendikartoteegi kaartide andmete struktureeritud eraldamiseks kasutatud OpenAI GPT-4o mudelit.[5] Kui teised meetodid välistati liigse ajakulu tõttu, siis selle, nagu ka teiste MLLM-ide puhul oli probleemiks mudelite ettearvamatus ja läbipaistmatus, mis avaldub nn andmehallutsinatsioonidena. See tähendab, et mudel annab enda arvates loogilise vastuse ka juhul, kui selleks puudub info, millele tugineda, tekitades seetõttu valesid ja olematuid fakte. Selliste juhtumite vältimiseks töötati välja praktika, mille käigus MLLM ise tähistab töö käigus kohad, kus tuvastamine võib raskusi valmistada, et inimesed pööraksid neile andmete läbivaatamisel suuremat tähelepanu.

---

2 https://www.transkribus.org/ (26.07.2025).

3 https://github.com/AI-Riksarkivet/htrflow

4 See massiivne info saab peatselt kasutajatele kättesaadavaks Rootsi Riigiarhiivi andmebaasi kaudu https://sok.riksarkivet.se/en/nad (26.07.2025).

5 https://svenskahistoriskapatent.se/

Suurte keelemudelite kasutamist võib konverentsi ettekannete põhjal pidada tugevaks trendiks, kuid kindlasti pole see saanud veel üldlevinud tavaks. Näiteks ei kasuta siinkirjutajad neid mudeleid Eesti vallakohtuprotokollide tekstikorpuse puhul, mis pärinevad peamiselt 19. sajandi teisest poolest ja on kirjutatud enamasti vanas kirjaviisis.

Gerth Jaanimäe andis oma ettekandes ülevaate sellest, kuidas ta kasutas vallakohtuprotokollide normaliseerimisel ehk tekstide vanalt kirjaviisilt tänapäevasele teisendamisel statistilist masintõlget, mis on suhteliselt vana võte ja kasutusel kõrvuti uuema BERT keelemudeliga. Statistiline masintõlge valib tõlke kõige tõenäolisemate sõna- ja fraasivastete põhjal, samal ajal kui BERT püüab analüüsida kogu lause tähendust ja konteksti korraga. BERT on nagu GPT-gi suur keelemudel, kuid erinevalt teksti genereerivast GPT-st, mis ennustab järgmise sõna ainult eelnevate sõnade põhjal, on BERT-i eesmärgiks teksti uurida olemasoleva teksti põhjal, vaadates mõlemal pool sõna asuvat konteksti. Meetodite kombineerimisel prooviti kindlaks teha, kas kõiki sõnu tekstis on üldse mõtet teisendada, vahest esinevad mõned sõnad tänapäevases kirjaviisis juba varasemalt.[6] Treeninghulga peal tehtud tulemused näitasid, et normaliseerimist vajavate sõnade osakaal eri murdealades varieerus vaid 32–51% vahel.

Normaliseerimise kvaliteet tõusis, kui statistilise masintõlke meetodil normaliseeriti vaid eelnevalt BERT-i abil välja sõelutud tänapäevasest kirjaviisist erinevaid sõnu. Võrreldes praeguste trendidega, kus kasutatakse peamiselt vaid suuri keelemudeleid, mõjub selline lähenemine vastuvoolu ujumisena. Siiski on oma kindel koht säilinud ka vanematel meetoditel, mis on sageli lihtsamad ja robustsemad, kuid suudavad väiksema andmehulgaga paremini hakkama saada.

Suhtlusvõrgustike analüüs ja visualiseerimine

Üks digihumanitaaria tuumakamaid arengusuundi on suhtlusvõrgustike analüüs, mis annab võimaluse uurida inimeste, instantside jms omavaheliste seoste avaldumist, mõtestada kultuurivõrgustikke, avastada muidu raskesti tuvastatavaid suhtemustreid jpm. Ka konverentsil tutvustati mitmesuguseid suhtlusvõrgustike analüüsivõimalusi. Paljud paneelis „Digitaalne ajalugu“ arutletud küsimused on aktuaalsed ka mitmele siinkirjutajale, näiteks kuidas ja milliseid andmeid koguda, kategoriseerida, võrrelda, omavahel siduda jms. Vahest kõige otsesemalt käsitlesid neid probleeme Aalto ülikooli arvutiteaduste osakonna professori Eero Hyvöneni ja doktorant Henna Poikkimäki ettekanded kirjavahetuse uurimisest (projekt „LetterSampo Finland 1809–1917“) .[7]

Hyvönen tutvustas interdistsiplinaarse uurimisrühma (arvutiteadlased, kultuuripärandi eksperdid, ajaloolased ja kunstiajaloolased) teekonda uue tööriista loomiseni,

---

[6] G. Jaanimäe. Improving the Accuracy of Normalizing Historical Estonian Texts by Combining Statistical Machine. DHNB2025 Conference Proceedings. https://journals.uio.no/dhnbpub/article/view/12299

[7] Selleks kasutatud Sampo on soomlaste välja töötatud mudel, mille tööriistu andmete analüüsiks ja visualiseerimiseks on rakendatud mitmesuguste projektide puhul. Näiteks on projekt „ParliamentSampo“ uurinud Soome parlamendis peetud kõnesid. https://seco.cs.aalto.fi/projects/semparl/en/ (26.07.2025).

milleks erinevad partnerid tegid pikka aega laialdast koostööd.[8] Projektis kasutati lisaks lähilugemisele suurandmete töötlemist ja suhtlusvõrgustiku analüüsi ning kaasati ka keeletehnoloogid, et hinnata kirjavahetuse põhjal nii erinevaid sotsiaalseid ja suhtlusvõrgustikke kui ka nende arenguid ja info levikut.

Vaadeldavasse projekti on koondatud üle 1,2 miljoni kirja metaandmed 13 andmekogust umbes 89 000 isiku kohta. Paraku tuleb suurte andmestike puhul silmitsi seista mitmete piirangutega, millest olulisem puudutab metaandmete kvaliteedi varieeruvust. Projektis ei töötatudki valdavalt mitte sadade tuhandete digiteeritud kirjade sisuga, vaid kirjade kogusid iseloomustavate arhiivifondide või säilikute kirjeldustega.

Tervikliku võrgustiku analüüsi tarvis on siiski vaja kasutada lõimitud ja tasakaalustatud andmestikku. Seetõttu on Hyvöneni arvates uuemate meetodite kõrval jätkuvalt olulisel kohal lähilugemine, uurijapoolne konteksti tundmine ja vastava andmestiku „lugemisoskus" tervikuna. Ka paljudest ettekannetest jäi kõlama tasakaalu otsimine automaatselt tehtavate tööetappide ja uurijapoolse käsitsi lähenemise vahel, et saavutada võimalikult sisukas ja kvaliteetne tulemus.

Suurte andmehulkade efekt avaldub laiemale kasutajaskonnale eelkõige visualiseerimise toel, mis toob esile seoseid, mustreid, arenguid, muutusi jpm ning esitab neist tõukuvaid uurimisküsimusi. Erinevad visualiseerimise vormid alates graafikutest ja sõlmedest kuni sõnavihmani tekitasid paratamatult küsimuse andmete kasutajate oskustest, mille võttis tabavalt kokku Mahendra Mahey ettekande pealkiri „Mida võiksid digitaalse kultuuripärandi mitmekesised kasutusviisid meile õpetada sotsiaalse arengu soodustamiseks kavandatavate tegevuste ja teenuste kohta?".

Mahey julgustas katsetama erinevate mäluasutuste andmekogudega ja kokku tooma erineva tausta ning IT-oskustega inimesi, et innustada neid uurima varju jäänud nähtuste uusi tahke ja viia need laiema publikuni. Digihumanitaaria üks olulisi väljakutseid on ka avalikkusega suhtlemise oskus. Digitaalne pärand on omandamas kasvavat rolli hariduses, samuti tajutakse selles ühiskondlikku potentsiaali – üha suurem osa kultuuripärandist on ligipääsetav veebi vahendusel, mis võimaldab uurijatel sellega tegeleda ajast ja geograafilisest kaugusest olenemata. Samuti rikastavad meie teadmisi ning loovad tugevamat sidet pärandiga ühisloome projektid, millesse igaüks saab anda oma panuse. Sarnast mõttekäiku väljendati teisteski kultuuripärandile keskendunud ettekannetes.

Seega võib kokkuvõtteks sedastada, et DHNB 2025 konverents andis põhjaliku ülevaate digihumanitaaria projektide aktuaalsetest teemadest, olulistest edusammudest ja kitsaskohtadest ning pakkus väärtusliku võimaluse kuulata uurijate kogemusi, ammutada inspiratsiooni, luua rahvusvahelisi kontakte ning hoida end kursis uute metoodiliste ja tehnoloogiliste arengutega.

Konverentsil osalemine ja ülevaate koostamine on seotud projektiga EKKD-TA10 „Infoeraldus ajalooliste institutsioonide protokollide (1880–1940) näitel".

---

[8] E. Hyvönen, P. Leskinen and J. Tuominen. LetterSampo. Historical Letters on the Semantic Web: A Framework and Its Application to Publishing and Using Epistolary Data of the Republic of Letters. – Journal on Computing and Cultural Heritage 2023, 16 (1), lk 1–23.