

Paberilt päringuks, trükisest andmebaasiks: „Album academicumi“ (1632–1710) struktureerimine, analüüs ja visualiseerimine tehisintellekti kaasabil¹

MEELIS FRIENDENTHAL

Sissejuhatus

Käesolev mõneti tehniline artikkel tõukub soovist dokumenteerida uurimistöö käigus välja arendatud tööprotsessi. Nimelt tekkis korduva kasutamise käigus vajadus saada paremat ülevaadet Eesti haridusajaloo olulise allika, Arvo Teringu koostatud teose „Album academicum der Universität Dorpat (Tartu) 1632–1710“ kohta ja analüüsida sealseid andmeid.² Eesmärgiks sai võetud muuta trükitud raamatu info struktureeritud ja analüüsitavaks digitaalseks andmebaasiks, võttes teatavaks eeskujuks projektid Repertorium Academicum Germanicum (RAG) ja Repertorium Academicum (REPAC).³

Teringu raamat hõlmab varauusaegse Tartu ülikooli kahte esimest tegevusjärku – Academia Gustaviana (1632–65) ja Academia Gustavo-Carolina (1690–1710) perioodi, mida eraldab Vene-Rootsi sõjast tingitud paus, mil ülikooli tegevus viidi ajutiselt Tallinna. Raamat sisaldab biograafilisi andmeid 1706 Tartu ülikooli sisse kantud üliõpilase kohta: sageli on sealjuures antud ka nende immatrikuleerimise täpne kuupäev, päritolu, sotsiaalne taust, kaitstud disputat-

¹ Artikli autor tänab retsensenti asjatundlike märkuste eest.

² Arvo Tering, *Album academicum der Universität Dorpat (Tartu) 1632–1710* (Tallinn: Valgus, 1984).

³ *Repertorium Academicum Germanicum*, <https://rag-online.org>, vaadatud 07.11.2025; *Repertorium Academicum*, <https://repac.ch>, vaadatud 07.11.2025.

sioonid, kirjutatud gratulatsioonid ja mõnikord ka õpitud teaduskonnad ja nende hilisem elukäik. Allika väärtusest hoolimata on selle formaat – trükitud raamat – takistuseks mustrite ja seoste otsimisel statistiliste meetoditega.

Trükitud ja käsikirjaliste (ajaloo)allikate ligipääsetavus ja analüüsivõimalused on viimasel ajal teinud läbi märkimisväärse muutuse. Digihumanitaaria meetodite ja tehisintellekti kasutamine ajalooliste andmete uurimisel on üsna laialt levinud.⁴ Siiski on palju materjali endiselt vaid digiteeritud ja esitatud nt PDF-kujul ja nende sisu jääb tihtipeale masinloetamatuks (või kui on masinloetavaks muudetud, siis sisaldab tekst sageli rohkesti otsingut takistavaid tekstituvastusvigu). Lisaks on niisugune info struktureerimata, mis piirab laiaulatuslike kvantitatiivsete ja kvalitatiivsete analüüside tegemist. Siin on keeruline leida ka ühte ning igas olukorras toimivat töövoogu, sest iga allikas on erinev ning nõuab seega niinimetatud personaalset lähenemist.

Projekti keskmes oli generatiivse tehisintellekti (edaspidi TI), täpsemalt Google'i Gemini keelemudelite perekonna (sealhulgas versioonid 2.0 ja 2.5) rakendamine. Selle tehnoloogia valik tugines nii isiklikele praktilistele katsetustele kui ka laiema hulga kasutajate hinnangutele. Projekti läbiviimise ajal näitas Eesti teadlaste loodud hindamisplatvorm „Tehisaru baromeeter“, et just Gemini-perekonna mudel on eesti keele oskuse poolest kõige võimekam⁵ ja ka vanade keelte puhul (ladina, vanakreeka, vana kirjaväisiga eesti ja läti keel) annab Gemini mudelite kasutamine praegu (2025 suvi) kõige paremaid tulemusi.

Artikli teine eesmärk on näidata, kuidas selline digiteerimine ja teksti masinloetavaks struktureerimine annab uusi visualiseerimise ja andmete töötlemise võimalusi. TI ei ole siin kindlasti mitte ajaloolise uurimistöö asendaja, vaid tööriist, mis võimaldab keskenduda keerukamatele analüütilistele ja tõlgenduslikele ülesannetele, automati-

⁴ Nt Daniel Little, „Artificial intelligence tools for historians“, <https://undsoc.org/2025/07/19/artificial-intelligence-tools-for-historians>, vaadatud 07.11.2025; Digihumanitaaria Keskus, <https://www.kirmus.ee/et/teadus/digihumanitaaria-keskus>, vaadatud 07.11.2025.

⁵ Tehisaru baromeeter on Tartu Ülikooli, Tallinna Tehnikaülikooli, Tallinna Ülikooli ja Eesti Keele Instituudi koostöös loodud platvorm, mis hindab ja järjestab keelemudeleid nende võimekuse alusel eesti keeles. Platvormi edetabeli kohaselt oli artikli kirjutamise ajal kõige kõrgema hinnanguga mudel Gemini 2.5 Pro. Vt lähemalt: *Tehisaru baromeeter*, <https://baromeeter.tartunlp.ai>, vaadatud 30.07.2025.

seerides aeganõudvat ja rutiinset andmetöötlust. See lähenemine on otseses kooskõlas Arvo Teringu enda taotlustega, kes oma teose eesõnas seadis eesmärgi samuti varasemaid publikatsioone parandada ja andmeid süstematiseerida.⁶ Tering leidis, et kuni 4,2% Academia Gustaviana tudengite nimedest oli varem eksitavalt kirja pandud. Tema eesmärk oli luua täpsem ja usaldusväärsem allikpublikatsioon. Käesolev projekt on selle taotluse otsene jätk – eesmärk on luua baas, mis oleks ka tulevikus hõlpsasti täiendatav ja kasutatav.

Kasutades kombinatsiooni Pythoni skriptidest ning Gemini tehisintellekti mudelist, oli võimalik automatiseerida suur osa andmete digiteerimisest ja struktureerimisest, mis käsitsi tehes oleks nõudnud pikka aega (mitmeid kuid?) ja tüütut mitme inimese tööd. Tulemuseks on paindlik ja juba praeguses staadiumis täiesti kasutatav ning laiendatav uurimisvahend. Lähilugemist ja teksti intellektuaalselt läbiseedimist see muidugi ei asenda, aga võimaldab kiiremini läbi töötada andmehulki ja leida nii-öelda nõela heinakuhjast.

Lähteallikas ja projekti eesmärgid

Arvo Teringu „Album academicum“ on väga hea prosopograafiline allikas Academia Gustaviana ja Academia Gustavo-Carolina perioodi üliõpilaskonna uurimiseks. Viies andmed struktureeritud kujule, saab liikuda üksikute biograafiate uurimiselt kollektiivse biograafia juurde. Prosopograafia on ajalooline meetod, mis uurib ühiste tunnustega inimgruppide elukäike, et tuvastada laiemad sotsiaalsed mustrid ja seosed.⁷ See on eriti hea meetod ajaloooperioidide puhul, kus üksikisikute kohta on andmeid napilt, võimaldades teha järeldusi grupi kui terviku kaudu.

Kuigi Tering ise on teinud osa esmaseid analüüse, leidsin töötegemise käigus seal ka eksitusi (näiteks tudengite immatrikuleerumiste arv teatud aastatel osutus kontrollimisel Teringu toodud tabelis matrikli sissekannete arvust erinevaks).⁸ Lisaks tekkisid küsimused, millele trükiteksti baasil on keeruline vastata, nagu:

⁶ Tering, *Album academicum*, 5–6. Käesoleva artikli põhimõtted on Arvo Tering samuti heaks kiitnud (e-kiri autorile 13.08.2025).

⁷ Lawrence Stone, „Prosopography“, *Daedalus* 100, 1 (1971), 46–79.

⁸ Tering, *Album academicum*, 20–23.

- Millistest piirkondadest pärines enim üliõpilasi teatud perioodil?
- Kas üliõpilased kirjutasiid gratulatsioone⁹ oma piirkonna tudengitele, ehk kas jagunemine natsioonidesse¹⁰ paistab välja ka kirjallikes dokumentides?
- Mis aegadel toimus kõige rohkem immatrikuleerumisi?

Projekti peamine eesmärk tulenes praktilisest vajadusest muuta see rikkalik, kuid staatiline andmestik dünaamiliseks masinloetavaks andmebaasiks. Selleks sai paika pandud järgmised vaheeesmärgid:

1. Kvaliteetse digitaalse teksti loomine raamatu skaneeringutest.
2. Teksti struktureerimine loogilisteks üksusteks (üliõpilaste kirjeteks).
3. Iga kirje sees olevate andmepunktide (nimi, kuupäev, päritolu, karjäär jne) eraldamine ja formaliseerimine.
4. Andmete salvestamine andmebaasi, mis võimaldab edasist analüüsi ja valideerimist.

Tööprotsessi kirjeldus: andmete digiteerimine ja struktureerimine

Tööprotsess koosnes mitmest etapist, milles kombineeriti automatiseeritud skripte ja Gemini keelemudeli võimalusi. Projekti käigus loodud skriptid ja andmed on publitseeritud Githubi platvormil.¹¹ Skriptide kirjutamisel on kasutatud ulatuslikult Gemini 2.5 Pro abi.

Esimene etapp: skaneerimisest tekstifailini ja OCR-i kvaliteedi võrdlus

Teringu raamat oli juba skaneeritud, aga selle masinloetavaks muutmine oli eraldi probleem, sest andmebaasi loomisel oli vaja kindel olla, et nimed ja muu selline on kõik õigesti kirjutatud. Katsetati esiteks mitut lahendust: kommertstarkvara Abbyy FineReader ja avatud lähtekoodi-

⁹ Õnnitlusluuletused olid osa varauusaegse ülikooli kirjanduslikust tegevusest, enamasti kirjutatud disputatsiooni kaitsmise puhul. Vt nt Kristi Viiding *et al.* (koost.), *O Dorpat, urbs addictissima musis ...: valik 17. sajandi Tartu juhuluulet* (Tallinn: Eesti Keele Sihtasutus, 2007).

¹⁰ Päritolu alusel loodud tudengite kolleegiumid. Vt Sten Lindroth, *Svensk lärdomshistoria. Stormaktstiden* (Stockholm: P. A. Norstedt & Söners förlag, 1975), 40 jj.

¹¹ <https://github.com/meelisf/album-academicum>, vaadatud 07.11.2025.

ga Tesseract (läbi ocrmypdf liidese) ning Google'i Gemini Flashi mudelit.¹² Kvalitatiivne visuaalne võrdlus (nn *vibe* hinnang) näitas, et Gemini andis antud allika puhul – mis sisaldab ladina- ja saksakeelseid termineid ning rohkelt lühendeid – märgatavalt kvaliteetsema ja puhtama tulemuse kui traditsioonilised OCR-tööriistad.¹³ See vähendas oluliselt käsitsi tehtava parandustöö mahtu ja oli hea indikaator, et nüüdisaegsed keelemudelid suudavad ajalootekstidega küllalt hästi toime tulla.

Teine etapp: teksti esmane tükeldamine Pythoni skriptidega
Saadud tekstifail vajas loogilisteks osadeks jagamist. Selleks kasutasin mitut Pythoni skripti, mis jagasid kogu teksti esmalt sadade kaupa plokkideks, seejärel grupeerisid kirjed kuupäevade järgi ning lõpuks eraldasid iga tudengi kirje eraldi tekstifailiks. Selle etapi lõpuks oli kogu raamatu sisu jagatud 1706 väikeseks failiks, millest igauks sisaldas ühe üliõpilase andmeid.

Kolmas etapp: struktureeritud JSON-andmete genereerimine Gemini abil

Järgmine eesmärk oli muuta iga tudengi vabas vormis tekstikirjeldus struktureeritud JSON-objektiks. JSON (JavaScript Object Notation) on standardne andmevahetusformaad, mida toetavad kõik levinud programmeerimiskeeled (nagu Python) ja andmeanalüüsi tarkvarad. Seega saab loodud andmestikku importida erinevatesse andmebaasidesse (nagu projektis kasutatud MongoDB) ja kasutada tulevikus uute analüüsi- ning visualiseerimisvahenditega. Lisaks sobib JSON ajalooliste andmete talletamiseks eriti hästi tänu paindlikkusele – ei ole vajadust, et igal kirjel oleksid täpselt samad väljad ning lisaks toetab see hierarhilisi ehk pesastatud struktuure (näiteks võimaldab üliõpilase kirje juures kirjeldada infot tema päritolu, vanemate, külastatud koolide ja muu sellise kohta).

¹² Githubis fail `ocr-few-shot.py`.

¹³ Ajaloolise materjali OCR-küsimuses vt nt Benjamin Kiessling, *Multilingual Party Model for European Languages*, Zenodo, June 28, 2025, <https://zenodo.org/records/15764161>, vaadatud 07.11.2025; Emanuela Boros and Maud Ehrmann, „Investigating OCR-Sensitive Neurons to Improve Entity Recognition in Historical Documents“, <https://doi.org/10.48550/arXiv.2409.16934>, vaadatud 07.11.2025. Samas pole „Album academicumi“ tekst otseselt ajalooline, vaid oluline oli saada võimalikult täpne transkriptsioon, eriti numbritest.

JSON-objektide loomiseks sai antud Gemini keelemudelile detailne instruksioon ehk viip (*prompt*), mis koosnes ülesande kirjeldusest, kontekstilisest lisainfost (lühendite loend, soovitud JSON-struktuuri kirjeldus), näidetest (*few-shot learning*) ja teisendatavast tekstist. Seejärel käivitati skript, mis saatis iga üksiku tekstifaili koos detailse instruksiooniga Gemini API-le (programmeerimisliidesele, mis võimaldab interneti kaudu andmeid töötluks saata), ning salvestas tagasi saadud JSON-formaadis vastuse.

Neljas etapp: andmete laadimine andmebaasi ja valideerimine

Genereeritud JSON-failid laaditi NoSQL andmebaasi MongoDB. On oluline rõhutada, et TI ei ole eksimatu. Kuigi tulemused olid valdavalt küllalt täpsed, vajasisid need siiski üsna põhjalikku ülevaatus ja kohendamist. Andmebaasis oli aga neid vigu juba oluliselt lihtsam märgata ja parandada kui algses tekstifailis.

Hea näide on Arvo Teringu kompaktne notatsioon disputatsiooni kirjeldamisel. Kirje 257 puhul on Teringul nt mainitud „Disp. 26. 6. 1640, 2. 1644 (Joh. Erici)“ – akadeemilisele lugejale on ilmne, et sulgudes mainitud *praeses*¹⁴ kehtib mõlema komaga eraldatud disputatsiooni kohta. Gemini mudel, järgides siin antud juhul aga süntaktilist loogikat, seostas *praeses*’e nime aga ainult vahetult eelneva kuupäevaga (2. 1644), jättes esimese disputatsiooni ilma eesistuja andmeteta. See näitab, kuidas TI ei suuda alati tabada autori spetsiifilisi ruumisäästlikke konventsioone, mis eeldavad lugejalt konteksti tundmist. Just selliste süstemaatiliste vigade tuvastamiseks ja parandamiseks osutus kriitilise väärtusega otsus säilitada iga JSON-kirje juures ka algtekst (JSON failis ’raw_text’ väli), mis tagab andmete auditeeritavuse. Tulevikus on plaanis selliste mustrite leidmiseks ja korrigeerimiseks rakendada järeltöötluks skripte, mis parandavad taolised spetsiifilised tõlgendusvead kogu andmestikus. See mitmeetapiline töövoog – esmane automaatne struktureerimine, seejärel reeglipõhine valideerimine ja parandamine – on oluliselt

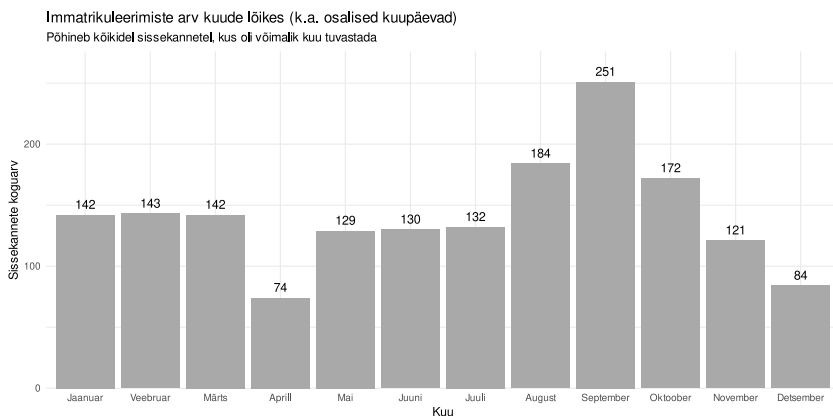
¹⁴ Disputatsiooni eesistuja. Varauusaegsetes väiketrukistes kasutatava terminoloogia kohta vt nt: Meelis Friedenthal, „Varauusaegsete disputatsioonide läbiviimisest, ülesehitusest ja teemadest“, *Keel ja Kirjandus* 64, 10 (2021), 865–885, <https://doi.org/10.54013/kk766a2>.

mõttekam kui katse luua eksimatut, kõiki erijuhte käsitlevat viipa. Lõpuks – ilmselt osa tööd tuleb teha ikka käsitsi, sest kõiki erijuhtumeid ei ole võimalik automatiseerida ning Gemini mudelile antava *few-shot*-õppe jaoks koguneb liiga palju erijuhtumeid, mis võib mudelit hoopis segadusse ajada. Kogemus näitab, et kui anda ette mitmeid väga erinevaid *few-shot*-näiteid, siis tekib mudelil teatav peataolek.

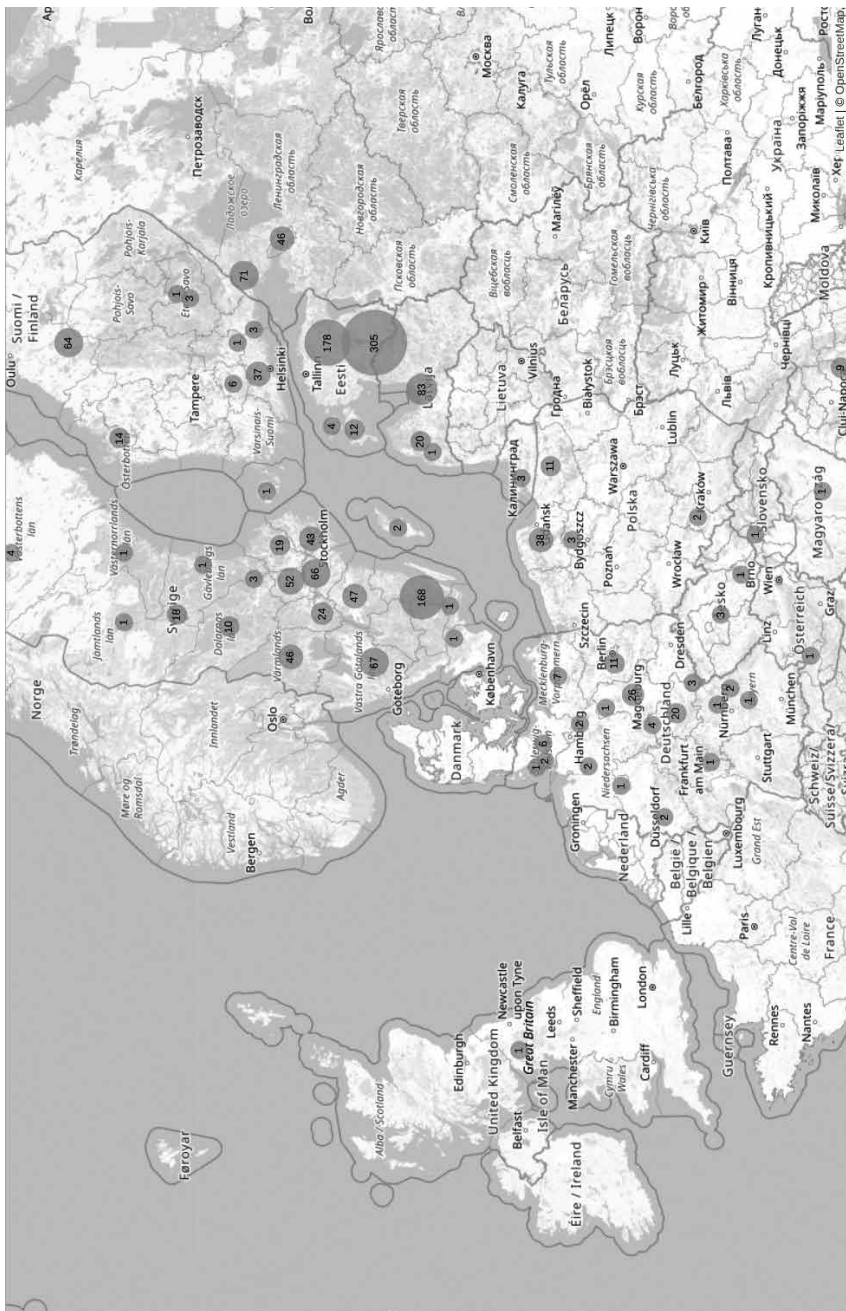
Teisisõnu leidis tõestust see, et TI genereeritud andmeid tuleb alati kontrollida ning inimene jääb ikkagi protsessi keskmesse kui ekspert ja lõplik vastutaja.

Andmestiku analüüs ja visualiseerimine: esmased tulemused

Selleks et demonstreerida loodud andmebaasi praktilist väärtust, sai läbi viidud R programmeerimiskeele abil (kasutades R-Studio) esmased statistilised analüüsid, mille eesmärgid oli visualiseerida üliõpilaste immatrikuleerumise jaotust kuude kaupa, geograafilist päritolu ja rahvusgruppide alusel koondumist. See protsess näitlikustab, kuidas struktureeritud andmed võimaldavad minna kaugele lihtsast tekstiotsingust.



Joonis 1. Immatrikuleerimiste arv 1632–1710 kuude kaupa (k.a. osalised kuupäevad). Põhineb kõikidel sissekannetel, kus oli võimalik kuu tuvastada.



Joonis 2. Academia Gustaviana ja Academia Gustavo-Carolina üliõpilaste päritolu piirkonnad perioodil 1632–1710.

Immatrikuleerumiste rütm

Joonis 1 iseloomustab varauusaegse ülikooli siseelu rütmi kogu perioodi vältel, näidates immatrikuleerimiste jaotumist kuude kaupa. Kuigi üliõpilasi võeti vastu igal ajal, on selgelt näha kaks peamist vastuvõtuhooaega. Kõige rohkem üliõpilasi saabus ja registreerus sügissemestri alguses: augustis, septembris ja oktoobris. Teine, mõnevõrra väiksem tipphetk on aasta alguses, jaanuarist märtsini, mis viitab kevadsemestri algusele. Väiksem vastuvõtt aasta lõpus, aprillis ja suvekuudel näitab, et akadeemiline aasta oli struktureeritud ning uute üliõpilaste saabumine koondus semestrite algusesse. See muster on omane paljudele varauusaegsetele ülikoolidele ja iseloomustab tollaseid sesooneid olusid (teede läbitavus, meresõiduks sobivad ajad ja muu selline) ning akadeemilisi praktikaid.

Päritolupiirkonnad

Esialgne trükitud allikast pärinev andmestik sisaldas väga varieeruvalt kirja pandud päritolukohti tekstilisel kujul (nt „Reval“, „Schles.“, „Tavastl.“). Geograafilise analüüsi tegemiseks oli vaja need kohanimed standardiseerida ja varustada koordinaatidega. Selleks loodi Pythoni skript, mis kasutas tasuta geograafilist andmebaasi¹⁵ GeoNames, et leida iga standardiseeritud kohanime jaoks geograafilised koordinaadid ja need andmebaasi lisada. Edasises töös on plaanis kasutada lisaks sellele täpsemaid ajaloolisi andmeid (näiteks Wikidata ajaliselt kvalifitseeritud andmed) ning dokumenteerida selgesõnaline normaliseerimistabel.

Struktureeritud andmestik võimaldab meil analüüsida ülikooli demograafilist arengut märksa detailsemalt, kui see oli võimalik varasemate meetoditega. Struktureeritud ja geokodeeritud andmete põhjal loodud interaktiivne kaart (joonis 2) pakub panoraamvaadet Academia Gustaviana ja Academia Gustavo-Carolina geograafilisest haardealast kogu perioodil 1632–1710. See visualiseerib ülikooli „mõjuala“ ehk peamisi värbamispiirkondi, kinnitades ja täpsustades kvantitatiivselt Arvo Teringu tekstilisi ja tabelites esitatud järeldusi.¹⁶ Iga ringi suurus kaardil on proportsionaalne sealt pärit üliõpilaste arvuga.

¹⁵ GeoNames, <https://www.geonames.org>, vaadatud 07.11.2025.

¹⁶ Vrd Teringu koostatud kaardiga „Album academicum“ sisekaanel: Tering, *Album academicum*.

Kaart visualiseerib tabavalt ka toleaegeid poliitilisi ja usulisi piire. Tähelepanuväärne on üliõpilaste peaaegu täielik puudumine Taanist, Norrast ja Poola-Leedu aladelt, mis olid Rootsi peamised poliitilised ja sõjalised rivaalid. Samuti näeme üksikuid „eksinud“ tudengeid kaugemalt, näiteks üks üliõpilane Inglismaalt ja mõned Kesk-Euroopast (sealhulgas Ungarist Siebenbürgenist ehk Transilvaaniast), mis viitavad ülikooli laiemale, ehkki ilmselt saksa keele-ruumiga piiratud tuntuks väljaspool Rootsi mõjusfääri.

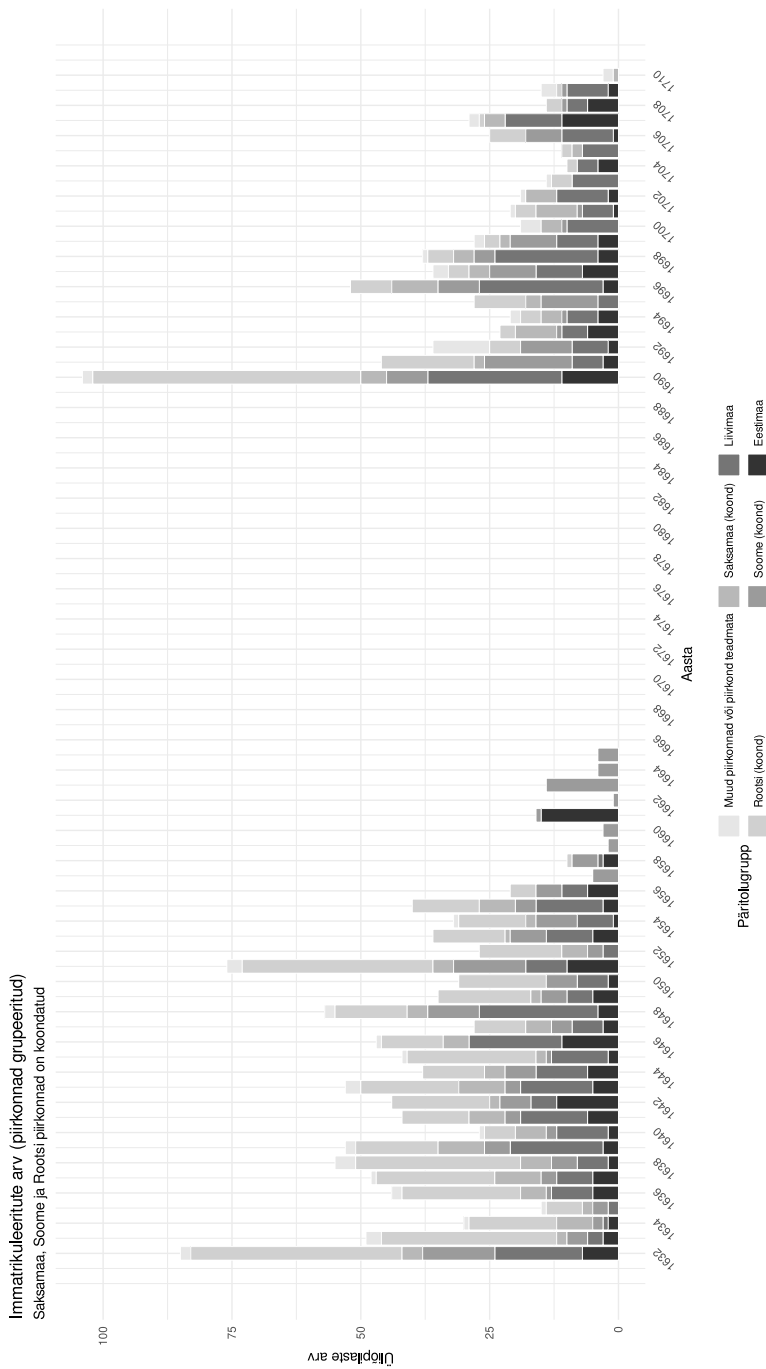
Kaardilt joonistuvad välja ka mitu selget kontsentratsiooniala, mis iseloomustavad ülikooli rolli ja tähtsust. Nagu oodatud, asuvad kõige suuremad üliõpilaste ringid ülikooli vahetus läheduses: Liivi-ja Eestimaal. See kinnitab ülikooli esmatahtsat rolli kohaliku, nii siinsete rootslaste kui ka baltisaksa eliidi olulise hariduskeskusena. Teine, peaaegu sama oluline värbamisala oli Rootsi emamaa. Eriti silmapaistvad on Lõuna-Rootsi piirkonnad, nagu Småland, Västra Götaland, ning tihe kobar üliõpilasi Stockholmi ja Uppsala ümbrusest. See olukord kajastab eelkõige Academia Gustaviana perioodi, mil ta suutis teatud määral konkureerida ka Rootsi suurriigi vanemate ülikoolidega (mida tõestab näiteks ka kaitstud disputatsioonide arv Tartu akadeemia esimesel perioodil).

Kaardil on selgelt näha Soome olulisus. Tähtis roll on siin Ingerimaal ja Karjalal, mis kinnitab Tartu ülikooli staatust Rootsi idaproovintside akadeemilise keskusena. Märkimisväärne tudengite kontsentratsioon on ka Põhja-Saksamaa aladelt, Rootsile kuulunud Pommerist, aga ka Poola mõjuväljas olnud Preisimaalt ja ka Hansa linnadest, nagu Hamburg. See peegeldab tihedaid poliitilisi, majanduslikke ja kultuurilisi sidemeid Läänemere regioonis ning akadeemilist mobiilsust luterlikus maailmas.

Immatrükuleerumised päritolu järgi

Kuhjatud tulpdiaagramm (joonis 3) visualiseerib immatrükuleeritud üliõpilaste koguarvu ja päritolu jaotust aastatel 1632–1710, pakku-des panoraamvaadet ülikooli haardeala muutustele läbi terve vaa-deldava perioodi.

Graafikult joonistuvad selgelt välja kaks peamist tegevusperioodi, mida eraldab Vene-Rootsi sõjast tingitud paus (1665–90) koos aastatega, mil tegutseti lühiajaliselt ja piiratud mahus Tallinnas



Joonis 3. 1632–1710 immatrikuleeritud üliõpilaste jagunemine päritolumaa järgi.

(väikesed tulbad pärast 1656. aastat). On tähelepanuväärne, et niinimetatud Tallinna-perioodil ei immatrikuleerunud Liivimaalt ega Rootsist pärit üliõpilasi.¹⁷

Mõlema selgelt eristuva perioodi alguses on näha suur immatrikuleeritute arv, kuid nende koosseis paljastab ülikooli arengus olulise muutuse. Ülikooli esimesel tegevusperioodil 1632–65 on näha Rootsi päritolu üliõpilaste (helehall) domineerimine. Avamisaastal 1632 moodustasid nad suure osa uutest tudengitest, mis kinnitab varasemat Teringu järeldust, et ülikool käivitati suuresti olemasoleva Tartu gümnaasiumi õpilaskonna ja Rootsist toodud tudengite baasil. Kuigi kohalikke, Eesti- ja Liivimaalt pärit üliõpilasi (tumedad toonid) lisandus pidevalt, jäi nende osakaal kohati rootslaste omale alla. Samuti on märgatav Soomest ja Saksa aladelt (hallid toonid) pärit tudengite püsiv, ent väiksem voog, mis kinnitab ülikooli haaret Läänemere ümbruses Rootsi suurriigi piires. See periood iseloomustab Tartu ülikooli kui Rootsi emamaale suunatud institutsiooni, mis oli paigutatud impeeriumi äärealale.

Ülikooli taasavamisel 1690. aastal näeme taas suurt immatrikuleeritute lainet, mis, nagu Teringu tekstist selgub, koosnes suures osas Uppsala ülikoolist Tartusse siirdunud kogenud tudengitest.¹⁸ Kõige olulisem muutus toimus aga järgmistel aastatel. Pärast esimest taasavamisjärgset lainet vähenes Rootsist pärit üliõpilaste osakaal drastiliselt ning esile tõusid Eesti- ja Liivimaalt pärit üliõpilased (tumedad toonid). Alates 1690. aastate keskpaigast moodustasid just kohalikud tudengid enamiku uutest immatrikuleeritutest. See näitab ülikooli rolli fundamentaalset muutumist: sellest oli saanud eelkõige kohaliku, Eesti- ja Liivimaa eliidi koolitamise keskus.

See tendents süvenes veelgi Põhjasõja aastatel pärast ülikooli üleviimist Pärnusse. Eriti tähelepanuväärne oli Soome piirkonna tudengite kadumine, ka Rootsi emamaa komponent kadus peaaegu täielikult, mis on ka mõistetav seoses sõjategevuse ja ühenduste ebakindlusega. Allesjäänud tudengkond oli valdavalt kohalikku pärit-

¹⁷ Tering, *Album academicum*, 20–21, 29 (viide 34). Tering grupeerib Eestimaa ja Liivimaa kubermangust pärit üliõpilased ühte gruppi „Livländer“. Siin analüüsis on lähtutud päritolukohast, nagu see on matrikliraamatusse kantud, tõlgendades seda vastavalt toleaegelele administratiivsele jaotusele.

¹⁸ Tering, *Album academicum*, 41.

olu, kuid samas tuleb arvestada, et teisel perioodil külastas Academia Gustavo-Carolinat ka rootsi soost tudengeid, kes märkisid oma päritolukohaks Liivi- või Eestimaa.

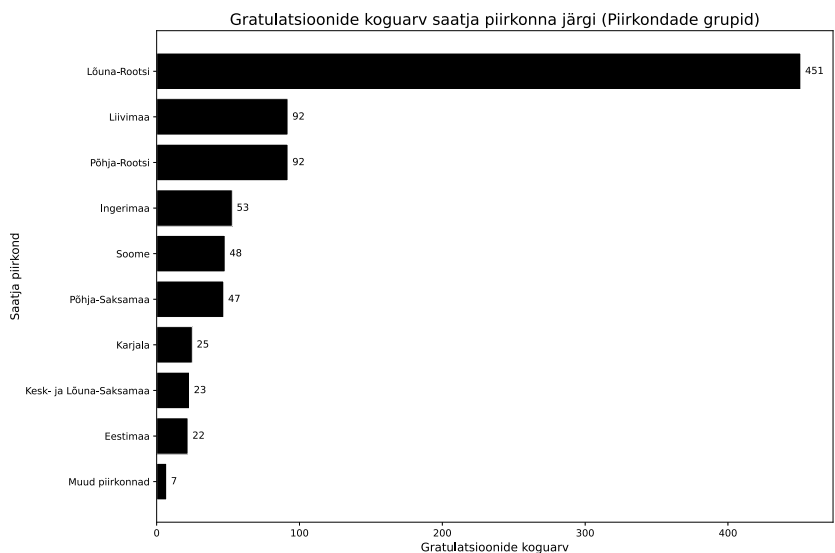
Kokkuvõtlikult eristuvad perioodid küllalt selgesti. Academia Gustaviana oli oma olemuselt eelkõige Rootsi suurriigi ülikool, mis asus Liivimaal, Academia Gustavo-Carolina muutus ajapikku üha enam provintsiülikooliks, mille peamine ülesanne oli harida kohalikku eliiti. Ilmselt oli Rootsi ja Soome päritolu õpilaste vähesus see, mis viis ka 17. sajandi lõpus pidevate kaebusteni, et ülikool ei ole hakanud loodetud aktiivsusega tööle. See visualisatsioon kinnitab ja nüansseerib Arvo Teringu järeldusi, pakkudes neile selge silmaga haaratava tõlgenduse.

Gratulatsioonid

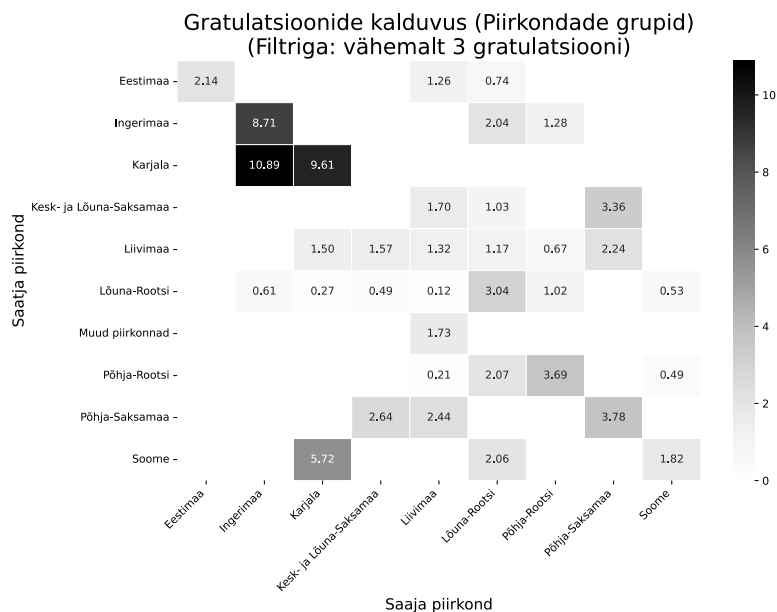
Lisaks staatilisele päritoluanalüüsile võimaldab struktureeritud andmestik uurida tudengite omavahelisi sotsiaalseid sidemeid. See oli üks peamisi küsimusi üldse andmestikuga töötama hakkamiseks. Üks häid indikaatoreid selliste seoste tuvastamisel on gratulatsioonid – akadeemilised õnnitlused, mida üliõpilased kirjutasid üksteisele disputatsioonides. Nimelt tuli Janika Pälli ja Jürgen Beyeriga vestlustest välja, et ei ole päris üheselt selge, kas tudengid kirjutasid gratulatsioone teistele tudengitele, kes parajasti samal ajal ülikoolis olid, või ilmneb siit ka regionaalne (rahvuslik) eelistus. Andmebaasi abil saab proovida sellele küsimusele statistika abil mõnesuguse vastuse leida – kes üldse gratulatsioone kirjutasid, kes keda õnnitles ja kas päritolupiirkondade põhjal joonistuvad välja kindlad kogukonnad?¹⁹

Lihtne gratulatsioonide arvu näitamine ei esita kogu pilti. Tulemusi mõjutab tugevalt asjaolu, kui palju üliõpilasi igast piirkonnast parajasti ülikoolis oli. Ilmselgelt olid kõige aktiivsemad selles vallas Rootsi päritolu tudengid (vt joonis 4). Selleks, et eristada teadlikku

¹⁹ Teringu „Album academicum“ on gratulatsioonide registreerimisel vähem täielik kui: Ene-Lille Jaanson, toim., *Tartu Ülikooli trükikoda 1632–1710: ajalugu ja trükiste bibliograafia, Druckerei der Universität Dorpat 1632–1710: Geschichte und Bibliographie der Druckschriften*, koostöös Mare Rand et al. (Tartu: Tartu Ülikooli raamatukogu, 2000). Samas puudub Jaansonil gratulatsioone kirjutanud üliõpilaste päritolu registreerimine. Edaspidi on plaanis ühendada Jaansoni ja Teringu andmed. Tänan Janika Pälli selle tähelepaneku eest.



Joonis 4. 1632–1710 kirjutatud gratulatsioonide arv saatja piirkonna järgi.



Joonis 5. Gratulatsioonide seosed gratulatsioonide saatja ja saaja vahel statistilise olulisuse järgi.

sotsiaalsed eelistust demograafilisest paratamatusest, tuli läbi viia kaheastmeline normaliseerimine.

Esiteks arvutati iga piirkondadevahelise seose jaoks kalduvuse (*propensity*) indeks. See indeks näitab, mitu korda rohkem (või vähem) õnnitlusi saadeti teatud gruppi võrreldes sellega, mida oleks võinud eeldada puhtalt juhusliku valiku alusel. Teiseks, et vältida väikestest valimitest tulenevat statistilist müra, rakendati tulemustele filter: visualiseeriti ainult need seosed, mis põhinesid vähemalt kolmel gratulatsioonil. Tulemuseks on soojuskaart (joonis 5), kus on esile toodud vaid statistiliselt kaalukamad eelistused või vältimised (kui andmeid ei ole või on liiga vähe, on selle asemel „auk“ ehk valge laik).

- Indeksi väärtus > 1 (tumedad toonid) näitab statistiliselt olulist eelistust.
- Indeksi väärtus ≈ 1 (hallid toonid) näitab neutraalset seost.
- Indeksi väärtus < 1 (helehallid toonid) näitab statistiliselt olulist vältimist.

Filtreeritud ja normaliseeritud andmed joonistavad välja üsna selge pildi ülikoolis eksisteerinud sotsiaalsetest võrgustikest ja barjääridest:

1. Ida-Soome ja Ingerimaa moodustasid tiheda kogukonna. Karjalast pärit tudengid eelistasid õnnitleda Ingerimaa kaaslast lausa 10,89 korda ja omasid 9,61 korda sagedamini, kui statistika eeldaks. Ka Ingerimaa-sisene eelistus oli erakordselt kõrge (8,71). See viitab väga tihedale kogukonnale, mis hoidis ülikoolis tugevalt kokku. Ka Soome tudengite Karjala eelistus oli märkimisväärne (5,72).
2. Normaliseerimine kinnitab, et rootslaste omavaheline suhtlus ei olnud juhus. Lõuna-Rootsi tudengid õnnitlesid omasid 3,04 korda ja Põhja-Rootsi tudengid 3,69 korda sagedamini, kui juhuslikkus eeldaks. Nende seoste aluseks on suur hulk andmeid, mis teeb järelduse küllalt kindlaks. Samavõrd kõnekad on helehallid toonid nende ridades: rootslased suhtlesid teiste gruppidega, eriti eesti- ja liivimaalastega (indeksid alla 0,3), märkimisväärselt vähem, kui oleks oodanud. See kinnitab tugeva sotsiaalse eelistuse ja samaaegse eraldumise mustrit.
3. Ka teised seosed muutuvad selgemaks – näiteks Kesk- ja Lõuna-Saksamaa tudengite side Põhja-Saksamaaga oli tugev (3,36), mis viitab ilmselt ühisele saksakeelsele kultuuriruumile.

Kokkuvõttes tõestab see mitmekihiline analüüs, et Tartu ülikooli tudengkond ei olnud homogeenne mass. Selle sees eksisteerisid selgelt piiritletud sotsiaalsed võrgustikud, mis ei põhinenud ainult päritolupiirkondade proportsioonidel, vaid ka aktiivsel sotsiaalsel eelistusel ja valikul.

Loodud andmestiku potentsiaal ja uued uurimisvõimalused

Struktureeritud andmebaasi loomine avab „Album academicumi“ uurimiseks mitmed uued võimalused, liikudes üksikute biograafiate uurimiselt kollektiivse biograafia ehk prosopograafia juurde.

- On võimalik uurida, kes kellega samal ajal õppis, millistest piirkondadest pärit tudengid moodustasid natsioone (*nationes*) ning kas päritolu ja õpingukaaslaste võrgustik mõjutas hilisemat karjääri.
- Andmebaasipäringud võimaldavad analüüsida ja visualiseerida, kuidas erineva sotsiaalse taustaga (näiteks pastori poeg *versus* kaupmehe poeg) üliõpilaste karjääriteed erinesid.
- Andmebaas võimaldab luua interaktiivseid kaarte üliõpilaste päritolukohtadest ja visualiseerida, kuidas see ajas muutus seoses sõdade, näljahäda või ülikooli maine muutustega.
- Digitaalset andmebaasi saab täiendada ja parandada lisaandmetega (näiteks professorite andmed, kirjavahetus, disputatsioonide andmed ja muu) oluliselt hõlpsamalt kui paberväljaannet.

Metoodilised järeldused ja tulevikuväljavaated

Läbi viidud uurimus kinnitab, et valdkonnapõhise konteksti ja näidetega abistatud tehisintellekt on küllalt hästi toimiv tööriist ajalooliste andmete struktureerimisel. Selle metoodika peamised eelised on kiirus ja niinimetatud skaleeritavus ning eriti oluline on eba-
korrapärase või ebatäieliku teksti töötlemise võimalus – selline tekst on aga trükitud ajalooallikate puhul pigem reegel kui erand.

Samas ilmnes andmete kontrollimisel lisaks juhuslikele teksti-
tuvastusvigadele ka (Gemini) TI tehtud süstemaatilisi tõlgendus-
vigu, mis on iseenesest oluline metodoloogiline tähelepanek. Need

eksimused ei viita niivõrd tehisintellekti puudulikkusele, kuivõrd illustreerivad ilmekalt inimese ja masina koostöö vältimatust ning vajadust anda tehisintellektile võimalikult täpseid ja läbimõeldud juhiseid ning viia läbi järelkontrolli ja parandusi.

Uurimuse järgmise sammuna on kavas andmeid täiendada kõigepealt Ene-Lille Jaansonite teoses „Tartu Ülikooli trükikoda 1632–1710“ sisalduvate materjalidega. Edaspidi tuleb kindlasti plaani võtta ka andmete ristseostamine teiste andmebaasidega, sidudes isikud ja kohad püsivate identifikaatoritega (nt Wikidata, VIAF). See annaks võimaluse andmestikku eksportida standardiseeritud vormingutesse, näiteks TEI-XML või RDF, ning liiguks seeläbi avatud linkandmete (*Linked Open Data*) põhimõtete rakendamise suunas. Ideaalis on võimalik luua ühendus sarnaste rahvusvaheliste projektidega.

Loodud metoodikat saab edaspidi rakendada ka paljude teiste allikate, näiteks senatiprotokollide, disputatsioonide, oratsioonide ja isikukirjavahetuse analüüsimisel ja seostamisel. Selliselt loodud andmekogud annaksid omakorda laiema platvormi Eesti ja kogu Läänemere regiooni varauusaja ajaloo uurimiseks, võimaldades mitmekesiseid otsinguid, statistilist analüüsi ja visualiseerimist.



Meelis Friedenthal, *PhD*, on varauusaja mõtteloo kaasprofessor ning Läänemere regiooni intellektuaalse ajaloo kaasprofessor.

From Print to Query, from Publication to Database: Structuring, Analysing and Visualising the *Album academicum* (1632–1710) with the Help of Artificial Intelligence

Meelis Friedenthal,
University of Tartu

This article discusses the process of converting the printed book *Album academicum der Universität Dorpat (Tartu) 1632–1710* – a list of students with short biographies from the early-modern university in Tartu/Pärnu compiled by Arvo Tering and published in 1984 – into a structured, analysable and visualisable digital database.

While digital humanities research is now well-established and the use of AI is increasing, much historical material remains in unstructured formats, such as PDFs, that are not suitable for large-scale analyses. By creating a database from the *Album academicum*, which is in a format very suitable for this process, this project represents a small contribution to filling this gap. The project's methodology was based on the combination of Google's Gemini family of large language models, which were chosen for their proficiency in Estonian and historical languages (here, Latin and German), and Python scripts. The process followed the spirit of Arvo Tering's original work, which was intended to correct and systematise earlier publications.

The article describes the project's multi-stage workflow. First, the printed book was digitised. Since it is significantly more capable at text recognition (OCR) than pre-existing tools such as ABBYY FineReader and Tesseract, using Gemini greatly reduced the need for manual correction. Second, Python scripts were used to segment the digitised text into 1,706 individual entries, one for each student. Third, a detailed prompt, including examples and a list of abbreviations, was provided to the Gemini API to convert each unstructured text entry into a structured JSON object (a few-shot learning approach). Finally, these JSON files were loaded into a MongoDB database for validation and manual correction. The article emphasises (with examples) that AI is not infallible and human oversight is crucial.

To demonstrate the database's potential utility, the article presents several initial analyses and visualisations.

The first analysis is of the rhythm of the academic year. It shows clear peaks in student matriculation in the autumn (August–October) and spring (January–March) semesters.

By standardising and geocoding students' places of origin using the GeoNames database, the project makes it possible to carry out spatial analyses. In this light, an interactive map was produced to visualise the university's geographical reach. This confirmed that the university's core recruitment areas were the local territories of Estonia and Livonia, the Swedish mainland, the northern provinces of Finland and Ingria, and the southern Baltic coast (notably, Pomerania and Prussia). The map also clearly illustrates the geopolitical boundaries of the era, with almost no students coming from rival powers such as Denmark and Poland-Lithuania.

A temporal analysis of student origins highlights a significant shift in the university's demographic profile over its two main periods of activity. In the period of 1632–56, during which the university was named *Academia Gustaviana*, the university was an imperial university, with the greatest number of students coming from Sweden proper. In contrast, over the period following its re-establishment under the name *Academia Gustavo-Carolina* (1690–1710), the university evolved into a provincial university that primarily served the local elites of Estonia and Livonia – a process that was significantly influenced by the Great Northern War, which discouraged foreign students from enrolling in the university.

Finally, the article presents a social network analysis based on academic gratulations between students. To move beyond raw counts, the data was normalised to calculate a „propensity“ index. This index shows which groups congratulated each other more or less often than would be expected by random chance. The results indicate the existence of distinct, tightly-knit social communities. For instance, an exceptionally strong bond was found between students from Karelia and Ingria. The analysis also confirms that students from Sweden showed a strong preference for congratulating other Swedes while interacting significantly less with students from Estonia and Livonia. These findings demonstrate that the student body was not a homo-

geneous entity but was composed of clearly defined social networks based on origin.

The article concludes that this project successfully demonstrates how modern AI models can provide powerful tools for structuring historical data, making it possible to transform static printed sources into dynamic research instruments. The resulting database (with necessary additions) opens up a new avenue for prosopographical research into topics such as social mobility, career paths and the formation of networks. The methodological findings underscore the importance of implementing a collaborative human-AI workflow, in which the AI automates routine data processing and the human expert provides crucial validation and contextual interpretation. Future plans include linking the dataset to other databases, such as Wikidata, and exporting the data into standardised formats (e.g. TEI-XML and RDF) to integrate it into the Linked Open Data ecosystem.