ESUKA – JEFUL 2019, 10–2: 25–44

DIGIDOKUMENDIST TEKSTIKORPUSEKS: SEMPERI JA BARBARUSE KIRJAVAHETUSE TÖÖTLEMINE MASINANALÜÜSITAVAKS PÄRINGUSÜSTEEMIS KORP

Marin Laak[1], Kaarel Veskis[1], Olga Gerassimenko[2], Neeme Kahusk[2] ja Kadri Vider[2]
[1]Eesti Kirjandusmuuseum ja [2]Tartu Ülikool

Kokkuvõte. Kirjandusteadlaste ja arvutilingvistide koostöös katseprojektina valminud Johannes Semperi ja Johannes Barbaruse kirjavahetuse korpus on nii kirjandusliku kui tekstilingvistiliselt huvipakkuv digitaalandmestik. Kirjandusteadlastele avab kaasaegsete digitaalsete meetodite kasutuselevõtt huvitavaid uurimisperspektiive ja vanade uurimistulemuste ülekontrollimise võimalusi arvutuslike meetoditega. Korpuslingvistidele on aga väljakutseks ajaloolise ja isikupärase keelekasutusega, erinevatest keeltest kubiseva ja rohkete koha-, aja- ja isikuviidetega tekstimaterjali ettevalmistamine rikkalikult märgendatud korpuseks. Artikkel peatub üksikasjalikumalt nii käsikirjalise materjali digitaalseks tekstiandmestikuks ettevalmistamise kui ka analüüsi- ja märgendamisprotsessi probleemidel ja nende võimalikel lahendustel. Kasutajatele tutvustatakse ka korpuste päringusüsteemi KORP võimalusi sarnaste tekstide uurimiseks.

Märksõnad: kirjandusteadus, erakirjad, digitaalne kultuuripärand, korpuslingvistika, loomuliku keele töötlus, andmekaeve, märgendamine

DOI: https://doi.org/10.12697/jeful.2019.10.2.02

1. Sissejuhatus

Tänapäeva uurijatena oleme jõudmas digiajastusse – järjest enam saavad uurimismaterjalid kättesaadavaks digitaalsel või digiteeritud kujul. Samamoodi nagu teistes Põhja- ja Baltimaades võib ka Eestis oodata digitaalse pärandi ja tekstiressursside järsku suurenemist: alanud on kultuuripärandi massdigiteerimise riiklik programm „Kultuuripärandi digiteerimine 2018–2023“ ja avaldatud selle

tegevuskava.[1] Uute digitaalsete ressursside loomine saab prioriteediks kõigile Eesti juhtivatele mäluasutustele; digiteeritakse eri tüüpi kultuuripärandit (trükitud raamatuid, arhiividokumente, foto- ja filmipärandit, esemeid ja kunsti). Digiteeritud ressursid tehakse kättesaadavaks internetis avaandmetena. (Tegevuskava 2018: 4) Kuigi paiguti jääb veel lahtiseks, kuidas ja mis eesmärgil neid digitaalseid ressursse kasutatakse (Laak, Viires 2015, 2016), toob digiressursside juurdekasv kaasa loomuliku keele töötluse võimalused ning teksti- ja andmekaeve meetodite kasvava olulisuse digihumanitaaria uurijatele, sest arvutuslike meetodite kasutamine muutub nii võimalikuks kui ka vajalikuks.

Milliste väljakutsetega me kokku puutume ja missuguseid teadmisi saame, kui hakkame kasutama keeletöötlusvahendeid digitaalse kirjandusliku pärandi uurimisel? Milliseid nõudeid esitab see digiandmete kirjeldamisele ja esitamisele, missugusi võimalusi pakub uurijale?

Käsikirjalise dokumendipärandi, nagu seda on ka kultuuri- ja kirjandusloolised arhiiviallikad, sh erakirjavahetused, digiteerimisel säilitatakse alusmaterjale üldjuhul liigendamata ja märgendamata tervikutena mõningate lisatud (arhiivinduslike) metaandmetega. Selle tulemusena on digiteeritud materjalid pigem inim- kui masinloetavad ning vastavad sellistena kirjandusteaduse metodoloogilistele traditsioonidele: töötada tekstide kui tervikutega, analüüsida nende poeetilisi ja semantilisi tähendusi jms. Traditsiooniliselt on kirjandusteadlased koolitatud kasutama peamiselt lähilugemise (ingl *close reading*) meetodit, harjumuspäraselt jätkatakse sellega ka suurtest digiteeritud tekstimassiividest vajaliku teabe otsimisel, mis omakorda seab piirid andmemahule (Viires, Laak 2018). Kuigi need väited tuginevad eesti kaasaegse kirjandusteaduse jälgimisele, näitab ka rahvusvaheline praktika, et kirjandusteadlased on digitaalhumanitaaria, sh korpusepäringu praktikate kasutuselevõtul olnud aeglasemad (vt Schreibman jt 2015) kui folkloristid ja lingvistid, kelle jaoks korpusepõhine uurimus on professionaalne standard juba ligi 30 aastat (vt ka Laak, Viires 2019: 131–132).

1 Vt lähemalt „Kultuuripärandi digiteerimise tegevuskava“ <https://www.kul.ee/et/eesmargid-tegevused/kultuuriparandi-digiteerimise-tegevuskava>.

Keeletöötluse ning teksti- ja andmekaeve meetodid nõuavad, et kirjanduslike teoste sõnaline tekst oleks muudetud digitaalseks, s.t andmeteks, mida saab analüüsida ning töödelda arvutuslikke programme kasutades.[2] See eeldab ka uusi oskusi andmeanalüüsi meetodites, üldist empiirilise uurimuse objekti ümbermõtestamist kirjandusteaduses, kuid mis peamine – senistest erinevate, uut tüüpi, võibolla isegi ambitsioonikamate uurimisküsimuste püstitamist. Näiteks võimalus võrrelda tekstistrateegiaid, retoorilisi ja stilistilisi mustreid omaelulooliste, kirjanduslike, usuliste ja poliitiliste tekstide korpustes võib anda meile uusi sissevaateid põimuvatesse ideoloogia, retoorika ja identiteedi esitustesse. Üks automaatse tekstianalüüsi huvitavamaid trende digihumanitaarias on näiteks meelestatuse analüüs (ingl *sentiment analysis*), mis võimaldab mõõta meelestatust parlamendi debattides (Rheault jt 2016).

2. Erakirjad kirjandusloo allikana

Aastal 2018 alustasime Eesti Kirjandusmuuseumi ja Eesti Keeleressursside Keskuse koostöös interdistsiplinaarse projektiga, milles on saanud kokku kirjandusteadus ja korpuslingvistika. Projekt keskendub kultuuripärandi, eriti arhiiviallikate kasutamisele uurimistöös loomuliku keele töötluse ning teksti- ja andmekaeve meetodite rakendamise kaudu.

Meie pilootkorpuse empiiriliseks aluseks on Eesti Kirjandusmuuseumi Eesti Kultuuriloolises Arhiivis säilitatav kirjanike ja tõlkijate, hiljem ka ühiskonnategelaste Johannes Semperi (1892–1970) ja Johannes Barbaruse (kodanikunimega Vares, 1890–1946) kirjavahetuse käsikiri.[3] Kirjavahetus kestis katkematult 29 aastat, esimene kiri on dateeritud 10. veebruaril 1911, viimane 10. jaanuaril 1940. Tegemist on käsikirjaliste erakirjade haruldase koguga, sest säilinud on mõlemad

2 Rahvusvahelises teaduses on kirjanduslike tekstide uurimisel kasutatud selliseid programme nagu R, Stylo ja Gephi.

3 Kirjavahetust oli aastatel 1978–1983 ette valmistatud kirjastuses Eesti Raamat avaldamiseks tekstikriitiliste kommentaaridega akadeemlise trükiväljaandena; käsikirjaga töötas Paul Rummo, hiljem Abel Nagelmaa ja Peeter Olesk. Käsikirjalistest kirjadest olid valminud masinkirjakoopiad, mille sisestas aastatel 2000 ja 2015–2016 kirjandusmuuseumis arvutisse Joel Ilja.

kirjavahetuse pooled.[4] Semper ja Barbarus olid lapsepõlvesõbrad, pinginaabrid ja aatekaaslased, kes said kokku Pärnu Poeglaste Gümnaasiumis (1905–1910). Nad jagasid sarnaseid intellektuaalseid hoiakuid, seisukohti ja väärtusi ning nad mõlemad olid frankofiilid, kes mitme aastakümne vältel vahendasid prantsuse kirjandust eesti kultuuriruumi ja vastupidi, tutvustasid kaasaegset eesti kirjandust prantsuse keeles (vt Laak 2017a: 214–216, 2017b: 1336).

Kuigi kirjavahetuses leidub rikkalikult kaasaegse igapäevaelu detaile, seisneb selle suurim väärtus siiski Eesti Vabariigi kirjandus- ja kultuurielu dokumenteerimises n-ö seestpoolt. See oli loomise ja vaidluste aeg Eesti kirjanduselus (Raid 2002), mil asutati kõik olulisemad kirjanduslikud institutsioonid alates Eesti Kirjanikkude Liidust (1922) ja selle ajakirjast Looming (1923), Eesti Kultuurkapital (1925) ning alustati Kirjastus OÜ „Looduse“ romaanivõitlustega, mis kõik mõjutasid oluliselt kirjanduspilti. Järgnevast kümnendist pakub kirjavahetus aga eriti unikaalse pildi, sest vapside liikumise ja sellele järgnenud vaikiva ajastu ainsaks usutavaks dokumendiks jäävadki eraallikad. See on kümnend, mil suursündmusena korraldati eesti raamatu aasta (1935), kümnendi lõpul asutatu ajakirjad Varamu ja Akadeemia. Semperi ja Barbaruse kirjavahetus valgustab seestpoolt ka varasemate kirjanduslike rühmituste nagu Siuru ja Tarapita ideid ja tegevust.

Kirjavahetuse ajal, 1920.–1930. aastatel on mõlemad autorid kirjandusloos tuntud eesti kirjandusavangardi pioneeridena (vt Hennoste 2016: 276, 340–346), Semper prosaistina ja Barbarus luuletajana; 1930. aastatel oli Semper Loomingu peatoimetaja, Barbarus aga majanduslikult sõltumatu linnaarst Pärnus, mis andis talle hea distantsi sündmusi hinnata. Kirjavahetuse vältel avaldas Barbarus 17 luulekogu, Semper neli luulekogu, kaks novellikogu ja kaks romaani (vt ka Laak jt 2019: 285). Kirju kirjutasid nad Eestimaa eri paigust, aga ka reisidelt: mõlemad kirjanikud reisisid palju Euroopas ning ka teistes maailmajagudes. Kirjades on nad tutvustanud vastastikku nii Eesti- kui ka välisreisidel külastatatud paiku, aga ka kontserte, kunstinäitusi, teatrietetendusi ning kirjanduslikke üritusi, rääkimata teistest kirjanikest,

4 Kirjade originaalid on hoiul eri arhiivides: Barbaruse kirju Semperile säilitatakse Eesti Kirjandusmuuseumi Eesti Kultuuriloolises Arhiivis (EKM EKLA f. 188, m. 1–3). Semperi kirjad Barbarusele leiduvad Eesti Rahvusarhiivis Johannes Varese kui riikliku tähtsusega isiku personaalkogus (ERA, f. R 39, n. 1, s. 14).

külalistest ja kohtumistest kõnelemisest. Nad kirjutasid teineteisele ka oma tööst, igapäevaelust, tervisest, arstirohtudest ja arvukatest hobidest, nagu jahilkäigud, sportimine jm. Kirjade vahendusel toimus isegi käsikirjade toimetamine ja raamatute kirjastamine.

Semperi ja Barbaruse kirjavahetuse keskseks ajaraamiks ja kontekstiks on periood kahe maailmasõja vahelises Euroopas ning elu Eesti Vabariigis aastatel 1918–1939, pakkudes seega kirjandusteadlastele semantiliselt rikkalikku ja mitmemõõtmelist uurimisainest. Meid on inspireerinud perspektiiv kasutada masinloetavasse vormi viidud kultuuriloolisi allikaid ja mina-dokumente (kirju, päevikuid, elulugusid jms) ka kirjandusteaduse interdistsiplinaarsetel lähialadel, näiteks elulookirjutuse (ingl *life writing studies*, vt Kurvet-Käosaar, Hinrikus 2013) uurimises.

Kahe eesti kirjaniku kirjavahetus 29 aasta vältel on empiirilise materjalina äärmiselt rikas teemade ja võimaluste poolest. See võimaldab püstitada nii traditsioonilisi kui ka uusi uurimisküsimusi: subjektiivsuse, emotsioonide ja sentimendi faktorid, mõlema autori verbaalsed, poeetilised väljendusvahendid, avangardne keeleloome jms. Kirjavahetuse kirjanduslik ja ajalooline väärtus on unikaalne, selle uurimine arvutuslike meetoditega on nii kirjandusteaduse kui korpuslingvistika vaatepunktist teedrajavaks (Keshabyan Ivanova ja Almela 2012). Kirjavahetus koosneb 670 kirjast, mille kogumaht on üle 1100 lehekülje ja 310 000 tekstiühiku, sh üle 310 000 sõne.

3. KORP kui kirjandusteadlase tööriist

KORP on korpuspäringusüsteem, mis võimaldab leida konkordantse ning teha eri parameetritel põhinevat statistilist analüüsi eri viisil märgendatud korpustest, kasutades teksti metaandmeid (autor, väljaandmise aasta, tekstitüüp jne) ning keelelist märgendust (lausestamine ja sõnestamine, punktuatsioon, morfoloogia, süntaks ja semantika) (Borin et al. 2012). Tehniliselt on KORP veebiteenus, mis kasutab taustal (avatud lähtekoodiga) korpuste töötlemise vahendit MS Open Corpus Workbench (Hardie 2012).

KORP on loodud Göteborgi Ülikoolis Rootsi Keelepangas (Språkbanken)[5], see võimaldab keelekasutust uurida eri tasanditel

5 Vt lähemalt <https://spraakbanken.gu.se/korp/>.

märgendatud korpustest ning saada vastuseks lisaks tavapärasele konkordantsile ka statistilisi näitajad ja kollokatsioone. KORPi arendatakse lisaks Rootsile veel mitmes riigis: Soomes Kielipankki[6], Norras Giellatekno taristu saami keelte jaoks[7], Taanis KORP[8], Islandil Risamálheildin[9].

Eesti KORPi[10] arendab Eesti Keeleressursside Keskus[11]. KORPis kättesaadavad korpused koosnevad praegu rohkem kui 850 miljonist tekstiühikust. Peaaegu kõik eestikeelsed korpused on automaatselt morfoloogiliselt märgendatud ja ühestatud morfoloogia tasandil võimaldamaks nii vormi- kui lemmapõhist otsingut.

Lisaks keeleteaduslikel eesmärkidel lisatud korpustele, mis on Eesti KORPis praegu valdavas enamuses, oleme katseprojektina loonud ka kirjandusteadlaste uurimishuvidele vastava Semperi ja Barbaruse kirjavahetuse korpuse[12] (kirjavahetuse aastad 1911–1940, korpuse maht 311 000 sõnet ja 21 500 lauset). Käsikirjalised originaalid olid juba eelnevalt arvutisse ümber kirjutatud. KORPiga ühitamiseks tuli teisendada trükitekst masinloetavaks andmestikuks: lisasime sellele käsitsi metaandmed ning lisasime automaatse vormianalüüsi ja ühestamise Vabamorfi[13] töövahenditega Giellatekno sõnaliikide ja grammatiliste kategooriate süsteemis[14].

Valisime oma projekti korpusepäringu süsteemiks KORPi, kuna see on avatud lähtekoodiga, paindlik ja lihtsalt õpitav süsteem, mis võimaldab graafilist ülevaadet alamkorpuste päringutulemustest, hõlpsat liikumist konkordantslausete ja laiema konteksti vahel ning ka statistikatulemuste ja näitelausete vahel, võimalusi grupeerida statistikat kõigi korpuses märgendatud kategooriate alusel, suhtelise esinemissageduse automaatarvutusi (miljoni korpusesõne kohta). Näitelauseid ja statistikat saab eksportida CSV (*Comma Separated Values*) ja JSon

6 Vt lähemalt <https://korp.csc.fi/>.
7 Vt lähemalt <http://gtweb.uit.no/korp/>.
8 Vt lähemalt <https://alf.hum.ku.dk/korp/>.
9 Vt lähemalt <http://malheildir.arnastofnun.is/>.
10 Vt lähemalt <https://korp.keeleressursid.ee/>.
11 Vt lähemalt <https://www.keeleressursid.ee/et/>.
12 Vt lähemalt <http://doi.org/10.15155/9-00-0000-0000-0000-00190L>.
13 Vt lähemalt <https://github.com/Filosoft/vabamorf>.
14 Vt <https://estnltk.github.io/estnltk/1.4/tutorials/morf_tables.html> (vaadatud 27.08.2019).

failivormingus. KORPi päringuvastuses tsiteeritud tekstilõigud on lause või lõigu pikkused, nii ei riku KORP autoriõigust, ületamata lubatud tsitaadi mahtu (AutÕS§19[15]). Päringutulemustes välja toodud metaandmed võimaldavad väga täpselt määrata näitelause asukohta kirjavahetuses, vajadusel on võimalik tekitada link mujal hoitavatele terviktekstidele, et pöörduda tagasi algallikate juurde.

KORPi abiga saab analüüsida ja objektiivselt kontrollida või täiustada ka Semperi ja Barbaruse kirjavahetuse[16] kohta esitatud uurimisväiteid. Näiteks võiks uurida, kas kirjandusteadlase Abel Nagelmaa (kirjavahetuse trükiväljaande toimetaja aastatel 1982–1983,) Eesti Kirjandusmuuseumi Eesti Kultuuriloolises Arhiivis (EKM EKLA 2012/121) oleva kokkuvõtte need väited kehtivad:

1) Semperi ja Barbaruse kirjavahetus on subjektiivne ja emotsionaalne, kirjad peegeldavad autorite iseloomu ja meeleolu nende kirjutamise hetkel;
2) kirjad demonstreerivad autorite seisukohti Eestit ja Euroopat puudutavates küsimustes;
3) kirjade temaatika sisaldab igapäevaelu, tervist, hobisid, külaskäike ja külalisi, kirjandustööd, raamatuid ja lugemist, aga ka kirjanduslikku, majanduslikku ja poliitilist elu Eestis ja Euroopas.

4. Käsikirjalise korpuse märgendamise väljakutsed

Tekstilise kultuuripärandi digiteerimisel võib eristada mitmeid etappe. Mitmed sammud on vajalikud konvertimaks digitaalsete tekstide kollektsiooni morfoloogiliselt analüüsitud korpuseks ning varustamaks kogu korpust ja tekste eraldi vajalike metaandmetega. Olenevalt püstitatud eesmärgist ei ole alati vaja kõiki neid läbi käia, kuid selleks, et paremini mõista tekkivaid võimalusi, vaatleme neid lühidalt:

1) digitaalne koopia (skaneeritud või pildistatud dokument) võimaldab üle saada füüsilise entiteedi ajalisest ja ruumilisest piiratusest, jagada seda mitmele inimesele, säilitades algse dokumendi silmaga nähtavad omadused;

15 Vt <https://www.riigiteataja.ee/akt/119032019055> (vaadatud 27.08.2019).

16 Vt lähemalt <https://korp.keeleressursid.ee/?mode=correspondence#>.

2) digitaalne koopia tekstist – see võib olla kas käsitsi ümber trükitud või skaneeringust tehtud tärktuvastus ehk OCR (ingl *Optical Character Recognition*). Teksti ümberkirjutus ei pruugi säilitada kõiki algse dokumendi visuaalseid omadusi, kuid pildiga seotud OCR võib ses osas rikkam olla;
3) märgendatud tekst. Tekst võib olla märgendatud väga mitmel tasandil, alates viidetest visuaalile (ridade ja lehekülgede algused ja lõpud) kuni omaette metatekstini (uurija kommentaarid). Märgenduse eri kihid võimaldavad otsida tekstist erinevaid nähtusi, nii lingvistilisi kui ka intratekstuaalseid. Esimesel märgendustasandil tuleks säilitada tekstitüübi olulised struktuuriüksused (pealkirjad, alamosad, värsid, kirjavahetuse puhul kirjad). Kirjavahetuse korpuse digitaalses algtekstis on märgendatud need osad: kirja algus ja lõpp, saatmise aasta ja kuupäev (nii kirjas esineval kujul kui ka normaliseeritult: AAAA-KK-PP), kirja kataloginumber, liik (nt piltpostkaart), autor ja adressaat, kirja saatmise koht, kirjale lisatud metainfo (märkused, varasemate uurijate kommentaarid).

Selleks, et märgendatud tekst oleks ka sisuliselt masinloetav ja -otsitav kogu vormirikkuses, peaksid olema märgendatud vähemalt lemmad (sõnade algvormid) ja sõnaliigid. Sõnestatud tekstist saame teha päringut küll sõnavormide (või nende osade) kaupa, aga mitte lemmade kaupa. Lemmatiseeritud tekstist saame samuti otsida lemmade ehk algvormide kaupa, aga mitte morfoloogiliste tunnuste kaupa (nt kääne, arv). Kui tekst on seejärel morfoloogiliselt analüüsitud ja märgendatud, saame otsida morfoloogiliste tunnuste ja sõnaliigi põhjal. Morfoloogiline ehk vormimärgendus täiustab märkimisväärselt korpuse kasutatavust ja kvaliteeti.

Kuidas saavutada seda, et tekst oleks vajalikul tasandil õigesti märgendatud? Ilmselt saab parima tulemuse nii, et teksti märgendab käsitsi inimene. See on aga väga ajamahukas, mis paneb küsima, kas saaks lasta teha analüüs ja märgendus masinal.

Eesti keele jaoks on olemas automaatne morfoanalüsaator juba 1990. aastatest. Praegu ilmselt kasutatavaim on Vabamorf[17], Filosofti

17 Olemas on mitmeid automaatse vormianalüüsi ja ühestamise tarkvaralahendusi eesti keele jaoks. Oleme selles projektis kasutanud eesti keele morfoanalüsaatorit Vabamorf <https://github.com/Filosoft/vabamorf>.

loodud Estmorfi avatud lähtekoodiga edasiarendus. Käesoleva kirjavahetuse korpuse automaatsel märgendamisel olemegi kasutanud Vabamorfi koos oletaja, statistilise ühestaja ja pärisnimede leidmise lisakomponentidega. Oletaja aitab analüsaatoril leida märgendeid sõnadele, mida ei ole analüsaatori leksikonis; pärisnimede määraja eelistab pärisnimeanalüüsi, kui analüüs on ebamäärane ja sõna algab suure algustähega.

Selleks, et morfoanalüsaator töötaks, on vaja, et tekst oleks sõnestatud ja lausestatud, see tähendab, et iga sõne, kaasa arvatud kirjavahemärk, oleks omaette tühikutega eraldatud ja iga lause omaette real. Sõnestamise ja lausestamise kvaliteet paneb aluse morfoloogilise märgenduse kvaliteedile. Kui sõnestamise puhul ei ole väga palju valikuid (siiski, näiteks kuupäevade vahemikku 8.–10. on võimalik sõnestada päris mitmel moel ja ka sidekriipsuga ühendatud sõnavorm *Under-Adsonitele* on võimalik märgendada ühe liitnime või kahe eraldi nimena), siis lausestamise puhul on oluline teha vahet lühendit lõpetaval kirjavahemärgil ja lauset lõpetaval kirjavahemärgil. Efektiivne lühendite märkimine aitab tublisti kaasa mitte ainult lausestamisele, vaid ka sõnaliikide märgendamisele.

Kirjavahetuse korpuses on palju ebastandardseid lühendeid, mis on olnud vastastikku arusaadavad ja mida autorid ei ole pidanud vajalikuks lahti seletada. Need on väljakutseks automaatsele vormianalüüsile ja ühestamisele, seda nii lühendite endi kui ka nende naabersõnade (ingl *collocate*) puhul, mille analüüs toetub kontekstile (nt *is*. *Linde* tuleb analüüsida *isand Linde* ja *Linde* on seega pärisnimi, mitte punktijärgne lausealguseline mitmuse osastav sõnast *lind*, nagu automaatühestaja punkti ja sellele järgneva suure algustähe järgi hetkel märgendab).

Automaatse lausestamise hõlbustamiseks on koostatud sagedusloend sellistest kirjavahetuses sisalduvatest punktiga lõppevatest sõnadest, mille viimane täht ei ole suurtäht ja millele ei järgne suurtähega algav sõna. Korpuses rohkem kui 9 korda esinevad lühendid (tabel 1) on lisatud korpuse teisendamisel kasutatud lausestusprogrammi.

Tabel 1. Kirjavahetuses sagedamini kui 9 korda kasutatud lühendid

Sagedus	Lühend	Seletus
64	jne.	ja nii edasi
55	kr.	kroon
51	a.	aasta
50	eks.	eksemplar
47	nr.	number
28	kirjandusl.	kirjanduslik
25	s.o.	see on
25	mk.	mark
24	näit.	näiteks
15	lhk.	lehekülg
13	kirj.	kirjandus, kirjanik
12	mrk.	mark
12	geom.	lühendina teose „Geomeetriline inimene“ pealkiri
10	sellep.	sellepärast

5. Millest sõltub vormimärgenduse kvaliteet?

Meie teada ei ole automaatse morfoloogilise märgenduse täpsust väga põhjalikult uuritud. 2008. aasta andmetel on käsitsi morfoloogiliselt ühestatud korpuse[18] peal kontrollitud automaatse märgenduse täpsus 93–98% (Veskis, Liba 2008). Kui suur võiks olla vigade hulk selles 20. sajandi alguse erakirjavahetuses, sellele küsimusele võime esialgu vaid oletamisi vastata, sest artikli kirjutamise hetkel oleme käsitsi läbi vaadanud ja parandanud vaid 4,87% kogu korpuse automaatsest märgendustest (kirjavahetus aastatest 1920 ja 1921, kokku 15 142 sõnet). Parandatud on kas lemmat või sõnaliiki või mõlemat, kuid parandusi ei ole tehtud teiste grammatiliste tunnuste osas. Sel moel said paranduse 1058 sõnet, mis teeb veaprotsendiks 6,99%. Selles arvestuses on ka ühestamata analüüside ühestamine sees, kuid seda ei ole tehtud järjekindlalt.

18 Vt lähemalt DOI:10.15155/1-00-0000-0000-0000-00085L.

Veidi mõjutab ühestamise täpsust ka meetodi valik. Kasutatud statistiline ühestaja ei vaata laiemat konteksti ning seetõttu on näiteks pärisnime *Linde* 47 esinemisjuhust ainult 11 saanud algvormiks *Linde*, järgnevad *linne* (12), *lind* (6) ja mitmed muud, ka ühestamata variandid.

Üks automaatset märgendamist raskendav asjaolu seisneb selles, et Semperi ja Barbaruse erakirjad ei olnud mõeldud laiemale publikule ega kirjutatud ametlikus kirjakeeles. Sageli tundsid mõlemad autorid kirjades mainitud isikuid ja sündmusi väga hästi ega pruukinud neid põhjalikumalt kirjeldada, piirdudes tihtipeale vaid vihjega sündmustele või inimestele. Ka kasutasid sõbrad hulgaliselt lühendeid, mis on olnud teada vaid neile endile, kajastades autorite isiklikku konteksti ja idiolekti. Avangardikirjanikena eksperimenteerisid Semper, eriti aga Barbarus palju ka keelega, kasutades oma loodud sõnu ja oma aja uudissõnu, sõnavorme ja väljendeid.

Semper ja Barbarus propageerisid Eestis prantsuse kirjandust ja kultuuri. Mõlemad autorid töötasid tõlkijatena ja reisisid palju. Nende kirjades on hulgaliselt sõnu, fraase, lauseid ja pikemaid lõike muudes keeltes peale eesti keele (ladina, prantsuse, vene ja saksa keeles). See on väljakutse keele automaatse analüüsi vahenditele, mida on treenitud eesti kirjakeele jaoks. Analüsaator ei tuvasta muukeelseid fraase, vaid oletab, et tegemist on talle veel tundmatu eestikeelse sõnavormiga. Sama lugu on tegelikult ka literaatide isikupärase keelepruugiga, mis saab standardkeele jaoks programmeeritud automaatanalüüsi tööriistadega vale analüüsi. Näiteks kasutavad mõlemad kirjasaatjad *n*-lõpulist seesütleva käände vormi üsna süstemaatiliselt. Kui edaspidi on võimalik keelelise koodivahetuse kohti täpselt tuvastada, lemmatiseerida ja analüüsida, oleks see info kindlasti kasulik nii keele- kui kirjandusteadlastele.

6. Grammatikast sügavamale

Semperi ja Barbaruse kirjavahetuse korpuse loomisel on olnud kaks laiemat eesmärki. Korpuslingvistika perspektiivist oleme uurinud tekstikorpuse koostamisel erakirjade erijooni: millised raskused ilmnevad töös sellise korpuse allikmaterjaliga; kuidas digiteeritud kirjade kogu konvertida keeleressursiks, et uurida keelelisi nähtusi digihumanitaaria meetoditega nende kirjanduslikes ja kultuurikontekstides?

Lisaks grammatilistele tunnustele on võimalik märgendada ka teisi huvipakkuvaid tekstiosi. Näiteks erinevate nimeüksuste, geograafiliste asukohtade, lühendite, muukeelsete tekstiosade ja ajaväljendite leidmine hõlbustab kirjanduse uurijatel nii hüpoteeside püstitamist kui ka uurimisküsimustele vastamist. Joonisel 1 oleme näitlikult märgendanud ühes kirjalõigus leiduvad erinevad andmetüübid.

„Bifur"i kiri saabus Sinu kirjaga ühel päeval. Palutakse kohe luule üle artikkel ära saata ja kedagi paluda, kes teisi küsimusi (sur la vie en général) käsitaks. Neil olla tõlkija, nii siis võivat eesti keeleski kirjutada. Et mul artikkel juba valmis oli, siis saatsin ta täna minema. Kui Sul lusti midagi saata, siis läkita kohe, – ehk novelli tõlge (maksavad 50 ***fr.*** *leheküljest), ehk siis mahutavad neljandamasse* ***nr****-isse; ehk viskad proosa & teaatri ülegi artikli. Aadress : „Bifur", Éditions du Carrefour, 199 , boul . St.-Germain , Paris (VIe) (M–eur le rédacteur en chef Ribemont Dessaignes). Küsisin kirjas, kas nende tõlkija luuletisi tõlkida võiks, siis võiksime valiku teha, ehk Suitsi ilmuva antoloogia neile saata. „Bifur" harrastab küll rohkem proosat & informatsioonilaadilisi ülevaateid, nii siis vaevalt nad luule liimile lähevad, aga eks ole ju veel teisi žurnaale pääle „Bifur"i, kus avaldada saaks, kui aga tõlkija leiduks. Mis teeb see* ***E.K.L.*** *propagandakomitee? Kas peab viimaks nende liikmete seas propagandeerima hakkama? Visnapuu kirjutab "V.-Maas" Igori pötserduse puhul, et vaja propagandeerida, aga, kui ma ei eksi, oli ta ise selles komitees?*

Joonis 1. Kirja lõik erinevate andmetüüpidega (kiri Barbaruselt Semprile 08.11.1929). Tähistused: (a) pärisnimi: isik, organisatsioon, pealkiri; (b) pärisnimi: asukoht, aadress; (c) **lühend**; (d) muukeelne aines; (e) ajaväljend, (f) isikupärane keelepruuk.

Selleks, et leida tekstist sündmusi ja nendevahelisi seoseid, tuleb tuvastada näiteks ajaväljendid. Lisaks kuupäevadele, mis on oluliseks osaks meta-andmetest, sisaldub kirjades endiski ajaväljendeid. Ka neid väljendeid on võimalik tarkvara abil teisendada konkreetseteks kuupäevadeks, mis võimaldab sündmuste ajateljele paigutamist.

Käsikirjaliste kirjade kuupäeva formaat ise on suure variatiivsusega: *11. jaanuar 1927*, *19/XII.37*, *2.2.1934*, *Jõulu 3. pühal 1934*. Autorite valitud kuupäevaformaat on säilitatud märgendatud tekstiosana, kuid lisaks sellele on kõik kuupäevad normaliseeritud nii, et neid oleks võimalik otsida ja järjestada (formaadis AAAA-KK-PP). Tuvastamata

kuupäeva(osade) kohta oleme kasutanud kategooriat „määratlemata“. Edaspidi loodame tekstis esinevaid ajaväljendeid automaatselt tuvastada ja märgendada EstNLTK[19] Pythoni teegi ja Siim Orasmaa arendatud ajaväljendite märgendus tarkvara abil (Orasmaa 2014, Orasmaa jt 2016).

Kirjandusteadlastele pakuvad suurt huvi pärisnimed, mida kirjades mainitakse: pärisnimede kasutus kombineerituna metaandmetega (ajaperiood, autor) võib viia oluliste mustrite ja tendentside tuvastamiseni. Nimeüksusi saab märgendada EstNLTK teegi abiga, eristades isikuid (PER), organisatsioone (ORG) ja kohti (LOC).

Praegune automaatne morfoloogiline analüüs ei erista näiteks isikunimesid kohanimedest. Siiski annab pärisnimedeks määratud sõnade algvormide sagedusloend võimaluse isiku- ja kohanimede kaardistamist puudutavate uurimiseesmärkide jaoks eeltööd teha (vt tabel 2). Mõned algvormid on oletuslikud (nt *Koms Komsi* asemel). Pärisnimedeks analüüsitud sõnade hulgas on ka mõned lühendid või arhailisemad sõnavormid *Terv*, *Pääle*, *Sääl*. Selliste vigadega tasub arvestada hilisemas analüüsis või morfoanalüsaatori kohandamisel (võimalik on edaspidi analüsaatorile õpetada, et *pääle* ja *sääl* jt kahe *ä*-ga vormid on samad, mis *peale*, *seal* jms).

Kõrvutasime pärisnimede arvu käsitsi morfoloogiliselt ühestatud korpusega[20] ja leidsime, et üldarv on võrreldav. Kuigi pärisnimesid on kirjavahetuses oluliselt rohkem kui käsitsi ühestatud ilukirjanduskorpuses, langeb nende hulk infotekstide ja ajakirjandustekstide vahelisse vahemikku. Pärisnimede liiane pakkumine on märgatav enim kirjade alguses, kus adressaadi nimele eelneb omadussõna *Armas*, mida automaatne analüüs peab pärisnimeks tegeliku (kuigi harvaesineva) eesnime *Armas* alusel.

19 Vt lähemalt <https://github.com/estnltk/estnltk>.

20 Vt doi:10.15155/1-00-0000-0000-0000-00085L.

Tabel 2. Kirjavahetuse korpuse pärisnimedeks analüüsitud algvormide sagedusloendi algus

Sagedus	Pärisnimi
1612	Pärnu
1240	Asm
1094	Armas
1088	Tartu
722	Tallinn
622	Barbarus
504	Pariis
398	Looming
392	Tuglas
330	Eesti
284	Berliin
234	Barb
234	Alle
220	Visna
188	Nyy
186	Siuts
176	Hispaania
170	Tarapita
166	Visnapuu

Kirjandusteaduse perspektiivist oleme testinud korpuslingvistika meetodite rakendatavust kirjandusteaduslikele uurimisküsimustele vastamisel. Üks korpuse kasutuse näidetest on olnud kirjavahetuses mainitud pärisnimede indeksi verifitseerimine. Indeks loodi käsitsi rohkem ligi 40 aastat tagasi kirjavahetuse trükiväljande jaoks ning see loendab kirjavahetuses mainitud võõrkirjanikke (EKM EKLA reg 1912/121). Indeksi alusel on valminud diagramm (vt joonis 2), millelt näeme ka seda, et kogu kirjavahetuses enim mainitud välisautor on prantsuse kirjanik André Gide.

Joonis 2. Võõrkirjanike ja maade mainimised kirjavahetuses 1983. aastal valminud käsikirja pärisnimede indeksi põhjal. Ringi suurus vastab mainimiste arvule korpuses.[21]

KORPi päring võimaldab näha Gide'i mainimisi sisaldavaid lauseid ning nende mainimiste absoluutsagedust ja suhtelist sagedust korpuses, kuid mitte üksnes seda. KORP võimaldab organiseerida statistikat kõigi kategooriate järgi, mida on korpuses märgendatud, sh metaandmete kategooriate järgi. Selleks, et teada, kes autoritest ja millal on maininud André Gide'i, peame vaid lisama metaandmete kategooriaid („autor" ja „aasta") statistika kriteeriumitele. Ka teisi metaandmeid ja keelelisi kategooriaid võib kasutada statistikas, mille pealt on võimalik liikuda näitelausete ja laiendatud konteksti juurde.

Statistikast (vt joonist 3) näeme, et Gide'ist rääkis kirjades põhiliselt Semper aastatel 1926–1930. See suhestub hästi ka kirjandusloogа: 1928. aastal kaitses Semper Tartu Ülikoolis kirjandusteaduse magistritöö „André Gide'i stiili struktuur" ja sai teadusmagistri kraadi (MA) ning jätkas tööd ülikoolis esteetika ja stilistika lektorina.

21 Graaf on tekitatud, kasutades Javascripti teeki D3 autorite nimekirjast ja kirjade arvust.

total_rows 15

sõna	autor	*kuupäev*	Kokku
Gide'i	Semper	-	6,4 (2)
Gide'i	Semper	1926-14-13	3,2 (1)
Gide'i	Semper	1928-05-26	3,2 (1)
Gide'il	Barbarus	1928-05-26	3,2 (1)
Gide'i	Semper	1929-04-03	3,2 (1)
Gide	Semper	1929-05-06	3,2 (1)
Gide'i-raama…	Semper	1929-07-31	3,2 (1)
Gide'i	Semper	1929-10-29	3,2 (1)
Gide'	Semper	1929-11-10	3,2 (1)
Gide'i	Barbarus	1929-12-08	6,4 (2)
Gide	Semper	1930-01-10	3,2 (1)
Gide'i	Semper	1930-10-13	3,2 (1)
Gide'i	Barbarus	1930-10-19	3,2 (1)
Gide'i	Semper	1931-07-03	3,2 (1)
Σ	Σ	Σ	51,5 (16)

Joonis 3. André Gide'i mainimiste statistika KORPis, sorteeritud sõnavormi, autori ja kuupäeva järgi. Esinemissagedused on esitatud eeldatava suhtelise sagedusena miljonisõnelises korpuses, sulgudes on toodud tegelik esinemiskordade arv (absoluutsagedus) Semperi ja Barbaruse kirjavahetuse korpuses.

7. Praegune ja tulevikutöö

Sõnade ja lausete arv kirjavahetuse korpuses on liiga suur selle märgendamiseks käsitsi, ent korpuseosa käsitsi märgendamine oleks vajalik. Ilma käsitsi märgendatud materjalita oleks raske (kui mitte võimatu) hinnata automaatset märgendust, mis on treenitud tänapäeva tekstidel. Kavas on märgendada väike osa korpusest ka käsitsi, parandamaks automaatset analüüsi.

Nagu varem mainisime, on kumbki autor kasutanud kirjades lisaks eesti keelele erinevaid võõrkeeli. Raske on öelda, kui palju, sest eesti keele automaatse morfoloogilise analüsaatori oletaja pakkus arusaadavatel põhjustel neis kohtades palju liiaseid ja valesid analüüse. Võiks proovida, kas keelemääraja (ingl *language detector*) enne morfoloogilist analüüsi parandaks tulemust, kuigi võib eeldada, et võõrsõnafraaside rohkuse tingimustes genereeriks see strateegia liiga palju müra.

Tekstis leiduvad olulised üksused, näiteks ajaväljendid või nimeüksused, võivad tulevikus samuti saada märgendatud ning muutuda päringukategooriateks. Ka süntaktiline ja semantiline märgendamine on võimalik ning realiseeritav tulevikus. See võimaldaks meil näiteks otsida teatud sõnatähendusi, jättes kõrvale sama sõna esinemised teistes tähendustes.

8. Kokkuvõte

Arhiividokumentidel nagu kirjanike erakirjavahetus on nii suur kultuuriväärtus kui ka keele- ja kirjandusteaduslik väärtus. Kirjandusteaduses korpuslingvistika meetodite rakendamine arhiivallikatele eeldab materjali hoolikat ettevalmistust arvutisse sisestatud tekstist tekstikorpuseks. Kirjanikest sõprade Johannes Semperi ja Johannes Barbaruse kirjavahetuse (670 saadetist, üle 1100 lehekülje) tekstikorpus (ligikaudu 310 000 sõnet) võimaldab meil edasi minna uut tüüpi uurimisküsimustega, alustades mõlema autori kirjanduslike mõjutuste selgitamisest kuni sügavamate tähenduslike küsimuste esitamiseni.

KORPi mitmekesised päringuvõimalused võimaldavad ka objektiivselt, korpuslingvistilist andmeanalüüsi appi võttes, verifitseerida järeldusi, mis on siiani tehtud, kasutades kirjandusteaduse traditsioonilisi meetodeid, põhiliselt tekstide lähilugemist.

Tänuavaldus

Haridus- ja Teadusministeeriumi uurimisprojekt „Kirjanduse formaalsed ja informaalsed võrgustikud kultuuriloo allikate põhjal“ (IUT2-2), EL Euroopa Liidu Regionaalarengu Fondi kaudu Eesti uuringute Tippkeskus (TK145) ja program ASTRA (2014-2020.4.01.16-0026). Uute tekstikorpuste lisamist päringusüsteemi KORP on toetanud ERFi projekt 2014-2020.4.01.16-0134 „Eesti Keeleressursside Keskuse (EKRK) ühendatud sisuotsing“ tegevusest „Riikliku tähtsusega teaduse infrastruktuuri toetamine teekaardi alusel“.

Aadress:
Marin Laak ja Kaarel Veskis
Eesti Kirjandusmuuseum
Vanemuise 42
51003 Tartu, Eesti
E-post: marin.laak@kirmus.ee ja kaarel.veskis@kirmus.ee

Kadri Vider, Neeme Kahusk ja Olga Gerassimenko
Arvutiteaduse instituut
Tartu Ülikool
Narva mnt 18
51009 Tartu, Eesti
E-post: kadri.vider@ut.ee, neeme.kahusk@ut.ee ja
olga.gerassimenko@ut.ee

Kirjandus

Borin, Lars, Markus Forsberg, Johan Roxendal (2012) „Korp – the corpus infrastructure of Sprakbanken". In N. Calzolari, K. Choukri, T. Declerck, M.U. Dogan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds. *Proceedings of the eight international conference on language resources and evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey.

Hardie, Andrew (2012) „CQPweb combining power, exibility and usability in a corpus analysis tool". *International Journal of Corpus Linguistics* 17, 3, 380–409. https://doi.org/doi:10.1075/ijcl.17.3.04har.

Hennoste, Tiit (2016) *Eesti kirjanduskriitika avangard 20. sajandi algul. Hüpped modernismi poole*. I. (Heuremata: humanitaarteaduslikud monograafiad.) Tallinn ja Tartu: Tartu Ülikooli Kirjastus.

Keshabyan Ivanova, Irina and Angela Almela (2012) „A new approach to literature: corpus linguistics". *International Journal of English Studies* 12, 2, 1–199. Saadaval Internetis <https://www.researchgate.net/publication/258868257_A_New_Approach_to_Literature_Corpus_Linguistics>.Vaadatud 30.04.2019.

Kurvet-Käosaar, Leena ja Rutt Hinrikus (2013) „Omaelulookirjutus taasiseseisvumisest nullindateni". *Methis* 11, 97–114. https://doi.org/10.7592/methis.v8i11.1004

Laak, Marin (2017a) „Kirjanduse sõdurid Esimese maailmasõja aegu: Barbaruse ja Semperi kirjavahetus 1911–1917". *Tuna* 2, 111–124.

Laak, Marin (2017b) „Kirjad Siuru päevilt: väljavõte Barbaruse ja Semperi kirjavahetusest". *Looming* 9, 1335–1347.

Laak, Marin, Kaarel Veskis, Olga Gerassimenko, Neeme Kahusk ja Kadri Vider (2019) „Literary studies meet corpus linguistics: Estonian pilot project of private letters

in KORP“. In C. Navarretta, M. Agirrezabal, and B. Maegaard, eds. *DHN 2019: digital humanities in the Nordic countries. Proceedings of the digital humanities in the Nordic countries 4th conference Copenhagen, Denmark, March 5-8, 2019*, 283–294. Copenhagen: University of Copenhagen, Faculty of Humanities. Saadaval Internetis <ceur-ws.org/Vol-2364/>. Vaadatud 10.08.2019.

Laak, Marin ja Piret Viires (2016) „Digital culture as part of Estonian cultural space in 2004–2014: current state and forecasts“. Saadaval Internetis <https://www.kogu.ee/vana/wp-content/uploads/2015/06/EIA-ENG OK-1.pdf>. Vaadatud 30.04.2019.

Laak, Marin ja Piret Viires (2015) „Digitaalkultuur Eesti kultuuriruumi osana 2004–2014: hetkeseis ja tulevikuprognoos“. Rmt R. Vetik, toim. *Eesti Inimarengu Aruanne 2014/2015 „Lõksudest välja?“*, 226–236. Tallinn: Eesti Koostöö Kogu.

Laak, Marin ja Piret Viires (2019) „Kirjandus ja digitehnoloogiad“. *Methis* 23, 129–147.

Orasmaa, Siim (2014) „Towards an integration of syntactic and temporal annotations in Estonian“. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, eds. *Proceedings of the ninth international conference on language resources and evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland.

Orasmaa, Siim, Timo Petmanson, Aleksander Tkachenko, Sven Laur, and Heiki-Jaan Kaalep (2016) „EstNLTK – NLP toolkit for Estonian“. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, and S. Piperidis, eds. *Proceedings of the tenth international conference on language resources and evaluation (LREC 2016)*. European Language Resources Association (ELRA), Paris, France.

Raid, Katrin (2002) *Loomise lugu. Eesti aeg: Eesti Kirjanikkude Liit 1922–1940*. Tallinn: Eesti Kirjanike Liit.

Rheault, Ludovic, Kaspar Beelen, Christopher Cochrane, and Graeme Hirst (2016) „Measuring emotion in parliamentary debates with automated textual analysis“. *PLOS ONE* 11, 12, e0168843. https://doi.org/10.1371/journal.pone.0168843

Schreibman, Susan, Ray Siemens, and John Unsworth (2015) *A new companion to digital humanities*. Chichester: Wiley Blackwell. https://doi.org/10.1002/9781118680605.

Semper, Johannes (1929) *André Gide'i stiili struktuur. Avec un résumé en français: le style d'André Gide.* (Akadeemilise Kirjandusühingu Toimetised / Publications de la Socété universitaire de litterérature à Tartu, 7.) Tartu: Varrak.

Veskis, Kaarel ja Erkki Liba (2008) *Automatic tagger evaluation: NLP course assignment report*. Saadaval Internetis <https://entu.keeleressursid.ee/public-document/entity-7052>. Vaadatud 30.04.2019.

Viires, Piret ja Marin Laak (2018) „Digital humanities meet literary studies: challenges facing Estonian scholarship“. In E. Mäkelä, M. Tolonen, and J. Tuominen, eds. *Digital humanities in the Nordic countries. 3rd conference. Helsinki 7-9 March 2018*. Helsinki: Helsinki University.

Tegevuskava 2018 = *Kultuuripärandi digiteerimine 2018–2023 tegevuskava*. Tallinn: Kultuuriministeerium. Saadaval Internetis <https://www.kul.ee/sites/kulminn/files/kultuuriparandi_digiteerimine_2018-2023_tegevuskava_1.pdf>. Vaadatud 30.04.2019.

Abstract. Marin Laak, Kaarel Veskis, Kadri Vider, Neeme Kahusk, and Olga Gerassimenko: Turning from digital document to text corpus: conversion of correspondence between Semper and Barbarus to a machine-readable unit in KORP. The article describes a joined pilot project of literary scholars and language technologists that resulted in a correspondence corpus of Estonian avant-garde poets Johannes Semper and Johannes Barbarus. The corpus is an inspiring digital dataset both for literary and linguistic researches. Contemporary digital methods allow literary scholars to find new interesting research perspectives and to revise the old research results with computational methods. Corpus linguists can find interesting challenges in historically and personally unique language use of the correspondents, in multiple languages used for citations and language play, in multiple references to places, events and persons in the textual material that was transformed to an annotated corpus. The article describes the preparation of typed-in manuscript material for a digital dataset in detail, problems of annotation and analysis and their possible solutions. The reader will get an insight to the possibilities that corpus query system KORP offers for the research of similar textual material.

Keywords: cultural heritage, literary studies, private letters, corpus linguistics, natural language processing, text and data mining, annotation