EESTI VANA KIRJAKEELE KORPUS: MIS TEHTUD, MIS TEOKSIL

VALVE-LIIVI KINGISEPP, KÜLLI PRILLOP, KÜLLI HABICHT

Sissejuhatus

Kui tosinkond aastat tagasi tutvustas Tartu Ülikooli praegune üldkeeleteaduse professor akadeemik Haldur Õim eesti lugejale lingvistilisi korpusi kui keeleuurimise olulisi andmebaase ja tõstatas eesti keele korpuse loomise vajaduse,[1] ei osanud keegi ette näha, kui kiiresti sel alal edasi liigutakse. Juba paari aasta pärast esitasid TÜ eesti kirjakeele korpuse koostajad ülevaate tekstikorpuste teooriast, tekstide valiku põhimõtetest, tekstitüüpidest ja tekstivaliku tehnilistest protseduuridest, samuti tekstide märgendamisest ning selleks ajaks valminud kasutusprogrammidest.[2] Vahepeal on arenenud kiiresti korpuse loojate oskused, ülikoolis on võimalik õppida arvutilingvistika eriala, luuakse mitut tüüpi korpusi, nt murdekorpus, suulise kõne korpus jt. Korpuspõhiste andmeteta ei kujuta enamik tänaseid eesti lingviste süsteemset keeleuurimist ette.

Vana kirjakeele alal alustati elektroonilise tekstikogu loomisega 1995. aasta sügisel, kui professor Mati Erelti algatusel loodi TÜ eesti keele õppetooli juurde vana kirjakeele töörühm. Rühma juhiks sai dotsent V.-L. Kingisepp ning olulisima tööülesandena alustati vanimate eestikeelsete tekstide arvutikorpuse loomist, eesmärgiga koostada selle põhjal vana kirjakeele sõnaraamat.[3] Praegune arvutiajastu on vana kirjakeele uurijate tööd märkimisväärselt kergendanud. Tekstikorpustesse ja andmebaasidesse talletatud materjali saab uurimishuvist lähtuvalt märgendada ja töödelda ning suurest materjalihulgast kiiresti sobivaid valikuid teha – see säästab uurijate aega ja energiat.

Vana kirjakeele korpusesse on praeguseks jõutud koondada umbes pool miljonit eestikeelset sõnet: kõik kuni XVI sajandini kirjutatud eestikeelsed tekstid, Georg Mülleri jutlused (1600–1606), lõunaeestikeelne katoliiklik käsiraamat "Agenda Parva" (1622), nn Turu käsikiri (XVII sajandi esimene pool), Joachim Rossihniuse lõunaeestikeelsed kirikuraamatud (1632), Heinrich Stahli kirikliku käsiraamatu "Hand- vnd Hauszbuch" (1632–1638) neli osa, samuti tema kaheosaline jutlusteraamat "Leyen Spiegel" (1641–1649) ning Christoph Blume neli eestikeelset teost: "Matthæi Judicis kleines Corpus Doctrinæ" (1662), "Geistliche Wochen-Arbeit" (1666), "Geistliche Hohe Fäst-Tahgs Freude" (1667) ja "Geistliche Seelen-Ergötzung" (1667). H. Stahli ja C. Blume teostega töö veel käib, sest sisestatud tekst vajab lemmatiseerimist ja grammatilise infoga varustamist. Kogu praeguseks korrastatud materjal on kasutajatele kättesaadav töörühma Interneti-koduleheküljel aadressil *http://www.murre.ut.ee/vakkur*.

[1] H. Õim, Lingvistilised korpused keeleuurimises. – Keel ja Kirjandus 1991, nr 5, lk 257–267.

[2] T. Hennoste, K. Muischnek, H. Potter, T. Roosmaa, Tartu Ülikooli eesti kirjakeele korpus: ülevaade tehtust ja probleemidest. – Keel ja Kirjandus 1993, nr 10, lk 587–600.

[3] Vt V.-L. Kingisepp, Eesti leksikograafia aastal 1997. Vana kirjakeele sõnaraamatust. – Keel ja Kirjandus 1997, nr 12, lk 825–827.

Korpusetekstide põhjal on tänaseks valminud mitu sõnastikku: XVI sajandi eestikeelsete tekstide sõnastik[4], Georg Mülleri jutluste[5] ja J. Rossihniuse kirikuraamatute sõnastik[6] ning nn Turu käsikirja sõnastik[7]. Lisaks leksikale on elektroonilised tekstikogud võimaldanud ka grammatikanähtuste uurimist (K. Prillop on uurinud G. Mülleri verbivormistikku[8], L. Merirand *pidama*-verbi grammatikaliseerumist[9] ning K. Kõpp vana kirjakeele relatiivlauset[10]).

Praeguseks on seoses grammatikanähtuste arengu süsteemsema uurimise vajadusega päevakorrale tõusnud pikemat perioodi hõlmava valikkorpuse koostamine, sest kui jätkata täielikku vanade tekstide ükshaaval sisestamist, ei saaks kirjakeele arengu uurimine korpuse põhjal veel kümnegi aasta jooksul võimalikuks.

Valikkorpuse üldised eesmärgid võiks kokku võtta järgmiselt: 1) pakkuda materjali uurijatele, kes vajavad mõne keelenähtuse iseloomustamiseks eri aegadest või autoritelt pärit võrdlusmaterjali (läbilõikeline diakroonilise andmestiku kaasamine); 2) pakkuda materjali vana kirjakeele uurijatele (leksika ja grammatika uurimine, autorite ja tekstide omavaheline võrdlus, kirjakeele arengu süstemaatiline kirjeldamine).

Uurija huvist ja vajadustest sõltuvalt peaks olema võimalik jälgida grammatikanähtusi ja leksikat eri sajanditel, eri peamurretest lähtuvais tekstides, eri tekstiliikides ja eri autoritel ning ka vana kirjakeele nähtusi tänapäeva keelega võrrelda. Lemmatiseeritud, s.t algvormile viidud, ühestatud ja grammatiliselt märgendatud korpus säästaks uurijaid mahuka, sageli keeruka ja osaliselt ka raskesti kättesaadava materjali käsitsi kogumisest.

Praegu puuduvad meil elektrooniliselt üldkättesaadavad tekstid XVII sajandi lõpukümnenditest kuni XIX sajandi viimase kümnendini. See tühik takistab aga kirjakeele arengust süsteemse ülevaate saamist.

Kirjakeele ajaloo korpusi mujal maailmas

Kõige tuntum keeleajalooline korpus on nn Helsingi korpus: inglise keele ajaloo korpus, mis on koostatud Helsingi ülikoolis. Korpus sisaldab umbes 1,5 miljonit sõna inglise keele erinevatest arenguetappidest (*old* 'vanainglise', *middle* 'keskinglise', *early modern English* 'varane uusinglise'), s.o aastatest 750–1710. Iga põhiperiood on omakorda jagatud 100-aastasteks alamperioodideks. Esindatud on hulk žanre (nii luule kui ka proosa), regionaalseid erinevusi ja sotsiolingvistilisi faktoreid (sugu, vanus, haridus, sotsiaalne klass) – korpus sobib seetõttu väga hästi keele muutumise (millal ja kuidas) uurimiseks.[11]

[4] E. Ehasalu, K. Habicht, V.-L. Kingisepp, J. Peebo, Eesti keele vanimad tekstid ja sõnastik. Tartu Ülikooli eesti keele õppetooli toimetised 6. Tartu: Tartu Ülikool, 1997.

[5] K. Habicht, V.-L. Kingisepp, U. Pirso, K. Prillop, Georg Mülleri jutluste sõnastik. Tartu Ülikooli eesti keele õppetooli toimetised 12. Tartu: Tartu Ülikool, 2000.

[6] V.-L. Kingisepp, K. Habicht, K. Prillop, Joachim Rossihniuse kirikumanuaalide leksika. Tartu Ülikooli eesti keele õppetooli toimetised 22. Tartu: Tartu Ülikool, 2002.

[7] H. Tennasilm, Turu käsikirja leksika. Tartu, 2002. Bakalaureusetöö käsikiri TÜ eesti keele õppetoolis.

[8] K. Prillop, Georg Mülleri jutluste verbivormistik. Tartu, 2001. Magistritöö käsikiri TÜ eesti keele õppetoolis.

[9] L. Merirand, *pidama*-verbi grammatikaliseerunud kasutusest vanemas kirjakeeles. Tartu, 2003. Bakalaureusetöö käsikiri TÜ eesti keele õppetoolis.

[10] K. Kõpp, Interrogatiiv-relatiivpronoomenid eesti vanemas kirjakeeles. Tartu, 2001. Bakalaureusetöö käsikiri TÜ eesti keele õppetoolis; K. Kõpp, Relatiivlausest eesti vanemas kirjakeeles. – Vana kirjakeel ühendab. Artiklikogumik. Tartu Ülikooli eesti keele õppetooli toimetised 24. Tartu: Tartu Ülikool, 2003, lk 116–128.

[11] M. Kytö, Manual to the Diachronic Part of the Helsinki Corpus of English Texts. Coding Conventions and Lists of Source Texts, 1996. – *http://helmer.hit.uib.no/icame/hc/*.

Helsingi korpusest on tehtud huvitavaid edasiarendusi, näiteks ”Penn-Helsinki parsed corpus of middle English”.[12] Selles korpuses on igast tekstist kolm versiooni: tavaline tekst, sõnaliigimärgenditega tekst ja süntaktiliselt märgendatud tekst (süntaktiline kategooria pluss grammatiline funktsioon). Eesmärgiks polnud mitte analüüsi absoluutne korrektsus, vaid süntaktilise info automaatse leidmise hõlbustamine.

Tuntuim märgendamata tekstide kogu on nn Augustana arhiiv, kuhu on koondatud ladina-, kreeka-, saksa-, inglis-, itaalia- ja hispaaniakeelseid ilukirjandustekste.[13]

Saksamaalt võib nimetada Mannheimi ajaloolist korpust, mis hõlmab tekste ajavahemikust 1700–1945 ning sisaldab 2,5 miljonit sõnet. Mõeldud on see korpus peamiselt sõnavara arengu ja grammatika diakrooniliseks uurimiseks.[14]

Ajaloolise portugali keele korpusesse (Tycho Brahe korpus) on valitud nende autorite tekstid, kelle sünniaasta jääb vahemikku 1550–1850. Iga sajandi igast veerandist on arvesse võetud kaks autorit, kelle emakeel on portugali keel ja kes pole pikka aega Portugalist eemal viibinud. Iga lisatud tekst peaks sisaldama vähemalt 50 000 sõna. Korpus on kavas morfoloogiliselt ja süntaktiliselt märgendada.[15]

Soome vana kirjakeele korpus hõlmab tähtsamaid terviktekste kuni 1810. aastani, kokku umbes 3,2 miljonit sõnet.[16] XIX sajandi kirjakeelest on soomlastel omaette valikkorpus, kuhu on võetud tekste võimalikult paljudelt eri aladelt, kusjuures tähtsaimaiks on peetud uusi teemasid, nagu tehnika ja sport.[17]

Vana kirjakeele korpuse eripära

Vana kirjakeele korpuse koostamine on töömahukam kui tänapäeva kirjakeele korpuse loomine. Tekstid on gooti kirjas ja sageli nii viletsa trükitehnilise kvaliteediga, et neid ei õnnestu skannides sisestada. Lisaks gooti kirjale peavad korpuse tegijad tundma ka saksa keelt, sest palju varasemaid allikaid sisaldab saksa-eesti rööpteksti. Ortograafia erineb oluliselt tänapäevasest, nii et sisestatud tekstide kontrollimisel pole abi tänapäeva eesti keele automaatsest õigekirjakontrollijast, tekstid tuleb käsitsi täht-tähelt üle vaadata, et kindlustada materjali usaldusväärsus.

Keeruline on valikkorpuse representatiivsuse tagamine. Esiteks seetõttu, et vana kirjakeel ei ole üks ega ühtlane keelekuju: eri piirkondades tegutsenud autorid esindavad ka XVII sajandil juba tekkima hakkava kirikliku kirjakeele traditsiooni taustal omaette keelekujusid, mistõttu korpus peaks kajastama neid kõiki. Mida vanema tekstiga on tegemist, seda olulisem on autori enese loodud keelenorm. Selles suhtes peaks vana kirjakeele korpus sarnanema rohkem murdekorpusega: nii nagu murdekorpus annab teavet eri murdealade ja murrakute keelepruugi kohta, peaks vana kirjakeele korpus esindama Eesti erinevates piirkondades töötanud autorite keele eripära. Oluline on seegi, et korpus sisaldaks eri autoritelt piisava pikkusega tekste, nii

[12] *http://www.ling.upenn.edu/mideng/*.
[13] *http://www.fh-augsburg.de/~harsch/augustana.html*.
[14] *http://www.ids-mannheim.de/lexik/HistorischesKorpus/*.
[15] *http://www.ime.usp.br/~tycho/corpus/*.
[16] *http://www.kotus.fi/aineistot/vks_sahkoinenaineisto.shtml*.
[17] *http://www.kotus.fi/aineistot/1800/1800_sahkoisetaineistot.shtml*.

et korpusetekst võimaldaks autori keele kohta järeldusi teha ja võrrelda tema keeletarvitust teiste autorite omaga.

Ei saa rääkida mingisugusest keskmisest vanast kirjakeelest. Näiteks Heinrich Stahl on *tahtma*-verbiga modaalkonstruktsioonides eelistanud supiini (87% juhtudest), Georg Mülleril esineb samas tarindis alati infinitiiv. Kui korpuses oleks mõlemalt autorilt proportsionaalse pikkusega tekstilõik, oleks *ma*- ja *da*-tegevusnimesid võrdselt, kuid sellisest valimist järelduv väide, et XVII sajandil põhjaeesti autorid ei osanud või ei pidanud vajalikuks õiget vormi kasutada, oleks eksitav: teada olevalt varieerus infinitiivi ja supiini kasutamine vaid mõnel autoril, näiteks H. Stahlil ja C. Blumel, mitte aga kõigil.

Teiseks representatiivsuse tagamist raskendavaks asjaoluks on tekstide vähesus ja see, et kõik tekstid ei ole säilinud. Esimesest säilinud eestikeelsest raamatust – Wanradti-Koelli katekismusest – on alles ainult 11 katkendlikku lehte. Praeguseni on leidmata ka R. Beseleri katekismus (1549), F. Witte katekismus (1554), S. Blankenhageni jutlustekogu (1630-ndate aastate lõpp) jt allikaid, mis on oma kaasaegset kirjakeelt kindlasti oluliselt mõjutanud. Isegi kui selgub, et mõnel perioodil kirjutatud tekstid on piisavalt sarnased, et uurida selle aja keskmist keelt, peab arvestama, et korpus saab esindada ainult säilinud tekste, kuid tekstide säilimine ei pruugi olla seotud nende olulisuse, mõju ega levikuga.

Korpuse ajalised piirid

Korpuse alguse üle otsustamine küsimusi ei tekita, sest kõige varasemad eestikeelsed tekstid pärinevad XVI sajandist ning esimene meie ajani säilinud trükitekst, Wanradti-Koelli katekismus, 1535. aastast. Varasemaid üksiksõnu ja -fraase ning kohanimesid sisaldavaid allikaid ei ole ilmselt esmajoones grammatika uurimiseks mõeldud korpuses mõttekas arvestada. Kui kaugele aga võib ajaskaala teises pooles liikuda, selles ei ole uurijad täiel üksmeelel ja seetõttu pole ka traditsiooni, millest lähtuda. Terminit *vana kirjakeel* seostavad mõningad lingvistid tekstidega, mis on loodud enne XX sajandit. Nii pärinevad eesti kirjakeele korpuse vanimad tekstid XIX sajandi 90-ndatest aastatest ja järelikult arvatakse, et kõik varasem ongi vana kirjakeel, mitte uus kirjakeel, kuigi sedalaadi vastandavat terminit pole kasutama hakatud.

Korpuse ajaliste piiride määramisel tulevad arvesse põhiliselt kaks laiemat aspekti: 1) kirjaviis, ühtse kirjakeele olemasolu ja autorite murdetaust, 2) tekstide hulk ja temaatika. Eesti kirjakeele 475-aastase ajaloo vältel on rakendatud mitut kirjaviisi ehk ajalooliselt kujunenud õigekirjatava: 1) korrapäratu, 2) vana ja 3) uus kirjaviis.

Korrapäratu kirjaviisiga tekstid püsisid XVII sajandi lõpuni, mil B. G. Forseliuse, A. Virginiuse ja J. Hornungi tööde kaudu hakkasid levima vana kirjaviisi põhimõtted, mille fikseeris J. Hornungi grammatika 1693. a. Senine korrapäratu märkimisviis asendus süsteemipärasema ja lihtsamaga. Eesti keele kirjapanemisel jõuti lähemale rahvakeele häälduse, võõrast ja saksapärasest hakati loobuma. Selles kirjaviisis ilmusid kirikukirjanduse tähtteosed Uus Testament (1715) ja Piibel (1739) ning see püsis võrdlemisi ühtlasena kogu XVIII sajandi ja XIX sajandi esimesel poolelgi.

Uut kirjaviisi hakati juurutama XIX sajandi keskel (E. Ahrensi grammatika 1843. a, selle täiendatud trükk 1853. a). See ei juurdunud üleöö ja varieerus autoriti veel kuni 1870-ndate aastateni, kui uues kirjaviisis ilmus um-

bes kolmveerand trükistest. Alles 1880. aastaks oli uues kirjaviisis 90% eestikeelsetest trükistest.[18]

Korratus kirjaviisis tekstid tuleks kindlasti lemmatiseerida ja morfoloogiliselt märgendada, sest muidu ei ole võimalik neist täpseid päringuid teha. Isegi kogenud vana kirjakeele uurija ei suuda alati välja mõelda ühe sõna kõiki võimalikke ortograafilisi kujusid (nt Mülleril sõna *neitsi* esinemiskujusid *Neutzist*, *Neutzÿ*, *Neuwtzist*, *Neutzi*, *Neützist*, *Newtzit*, *Neutzit*, *Neüwtzit*, *Neutzide*). Vanas kirjaviisis tekstid võiksid samuti olla lemmatiseeritud ja morfoloogiliselt märgendatud, sest see kirjaviis põhjustab eri sõnade kirjapildi ootamatut kokkulangemist ja on kergesti loetav vaid neile, kes on selle põhimõtetega kursis. Uues kirjaviisis tekstide eelnev lemmatiseerimine pole aga tingimata vajalik: nende analüüsimiseks saab kasutada olemasolevaid morfoloogiaanalüsaatoreid ja ühestajaid, mis korrapäratu ega vana kirjaviisiga tekstide puhul ei ole piisavalt täpsed. Valikkorpus võiks aga olla algusest lõpuni tehtav samade põhimõtete järgi.

Kirjaviisi ühtlustamisega seondub ka ühtse normeeritud kirjakeele kujunemine. Arvestama peab aga seda, et kirjaviisi ühtlustamine ei tähendanud grammatika ühtlustamist. Veel XIX sajandi algul loodi tekste kahes põhimurdes. Suuremat võidukäiku alustas põhjaeestimurdeline kirjasõna eeskätt O. W. Masingu, F. G. Arveliuse, J. W. Luce, O. R. Holtzi jt loominguga. Siiski lisasid autorid oma kirjatöödesse ka kodukandi murdejooni. Ühe ühise tallinnapärase kirjakeele taotlustega tuli välja J. H. Rosenplänteri Beiträge, kuigi selles avaldati ka tartu keele eeliseid esiletoovaid kirjutisi (nt Steingrüberilt, Hellerilt). Kahes peamurdes ilmus trükiseid kuni rahvusliku liikumise kõrgajani, mil saavutas suurema mõjuvõimu tallinna kirjakeel (1860-ndatel aastatel väheneb lõunaeestikeelsete trükiste osakaal 5 %-ni trükiste arvust).

Korpuse ajapiiride üle otsustamisel võiks arvestada ka säilinud tekstide hulka. Joonis 1 kujutab allikate hinnangulist sõnede arvu kümnendite kaupa (kordustrükke, grammatikaid, aabitsaid ja luulet arvestamata).[19]

Sõnede hulk kahekordistub 1840-ndatel aastatel, s.o ajal, mil algab vana ja uue kirjaviisi vaheline võitlus, ilmalik kirjandus on saavutamas ülekaalu, mh ilmuvad esimesed F. R. Kreutzwaldi proosateosed. Vana kirjakeele korpuse lõpp-piiriks võikski seda muutust arvestades olla 1840. aasta.

Võttes aluseks eestikeelsete tekstide hulga ja nende spetsiifika, võiks kirjakeele ajaloo jagada lihtsustavalt järgmistesse perioodidesse:

1) vana kirjakeele periood 1535. aastast kuni 1840-ndate aastate alguseni (võõrapärane, varieeruv, valdavalt usukirjanduse keel);

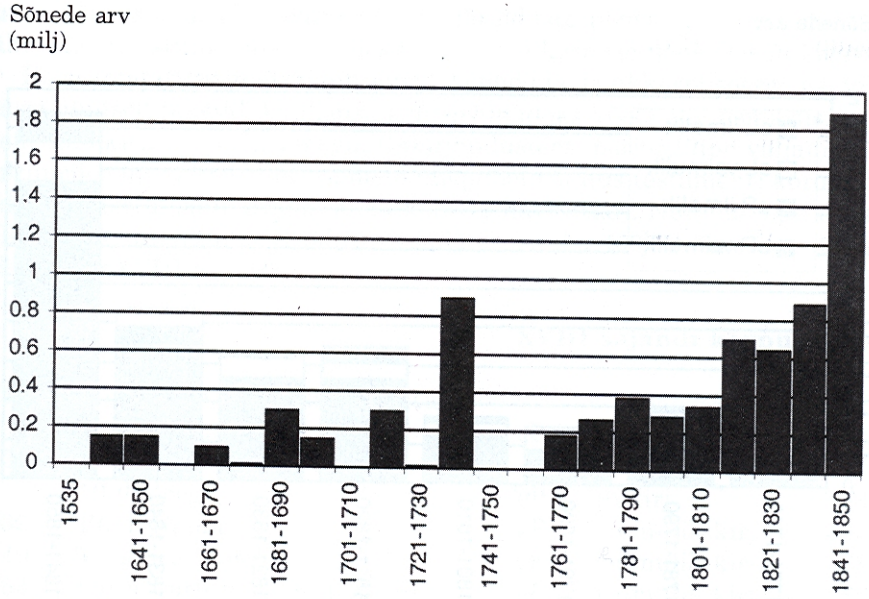
2) rahvapärastuva kirjakeele periood 1840-ndate aastate algusest kuni 1880-ndate aastateni (vaheetapp, mis algab kirjaviisi uuendamise katsetega: rahvusliku kirjakeele väärtustamine, eesti soost kirjameeste ja ilmaliku kirjanduse esiletõus, Eesti Kirjameeste Seltsi loomine);

3) normitud kultuurkeele periood 1880-ndatest aastatest kuni tänaseni (kirjakeele teadlik normimine kui pidev protsess, mille tulemusena on praeguseks välja kujundatud paljude allkeeltega eesti kultuurkeel).

---

[18] H. L a a n e k a s k, Kirjakeel. – Eesti entsüklopeedia 11. kd. Eesti. Üld. Tallinn: Eesti Entsüklopeediakirjastus, 2002, lk 593.

[19] Arvutuste aluseks olnud tekstivalik on tehtud "Eesti retrospektiivse rahvusbibliograafia" I osa põhjal (toimetanud E. Annus. Tallinn, 2000).



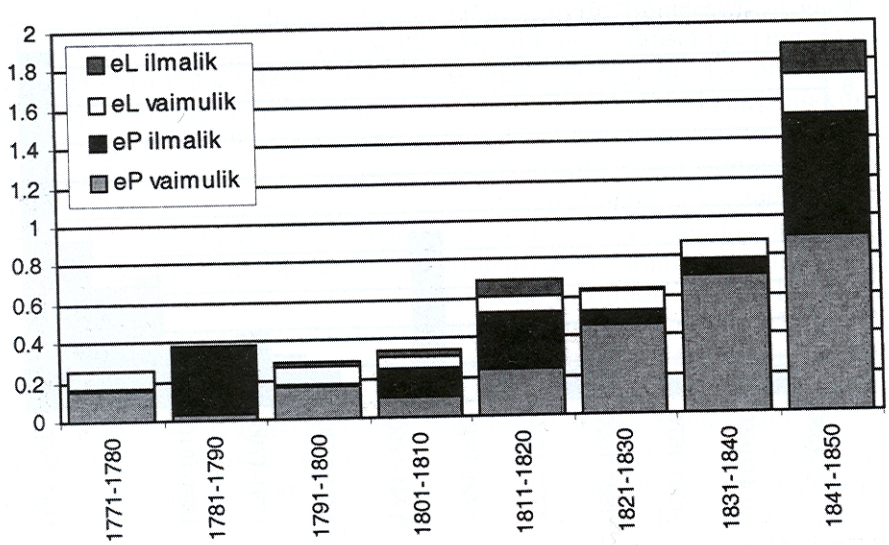
Joonis 1. Sõnede hinnanguline arv ajavahemikus 1535–1850.

Tekstide valikukriteeriumid

Korpuse koostamisel on olulisim probleem tekstivalik, s.t missugustelt autoritelt kui pikki tekstilõike kaasata. Tekstide pikkusest sõltub ka see, missuguseid keelenähtusi korpus uurida võimaldab. Vana kirjakeele valikkorpusest võiks välja jätta kordustrükid, riimilised tekstid, aabitsad ja kalendrid, üldjuhul ka käsikirjad. Kordustrükkide esmatrükkidega võrdlemine oleks töömahukuse tõttu mõeldamatu ega olegi üksikerinevuste tuvastamise korral mõttekas samale teosele korpuses liiga suurt osakaalu anda. Riimilised tekstid on allutatud vorminõudele ega anna seetõttu grammatikanähtuste kohta adekvaatset pilti. Vanemates aabitsates ei leidu originaalset teksti, need esitavad lisaks tähestiku tutvustamisele lugemise hõlbustamise eesmärgil silbitatud katekismuseteksti. Kalendrite puhul ei huvita meid keeleliselt kalendaariumi osa ning "kalendrisabades" esitatud ilukirjanduslooming on sageli kas varem või hiljem trükis ilmunud. Käsikirjalistest tekstidest tuleks teha erand vanimate tekstide osas, sest XVI sajandist ja XVII sajandi algusest on säilinud sedavõrd vähe tekste, et kõik teadaolev oleks mõttekas arvesse võtta. XVII sajandi teisest poolest ning XVIII ja XIX sajandist pärit käsikirjalisi tekste pole aga mõeldav korpusse kaasata nende väljaselgitamiseks kuluva töö mahu tõttu.

Tekstide valikul peaks tagama tartu ja tallinna keele ning ilmaliku ja vaimuliku kirjanduse piisava esindatuse. Joonisel 2 on kujutatud ilmaliku ja vaimuliku, tallinna- ja tartukeelse kirjanduse sõnede hulk kümnendite kaupa alates esimestest ilmalikest teostest.

Diagrammilt on näha, et aastatel 1781–1790 suureneb oluliselt tallinnakeelsete ilmalike tekstide osatähtsus, samal ajal kui tartukeelset kirjandust ei ilmu. Aastatel 1801–1810 on esindatud nii tartu- kui ka tallinnakeelne ilmalik ja vaimulik kirjandus. Tallinnakeelseid allikaid on küll rohkem, kuid ilmaliku ja vaimuliku kirjanduse proportsioon on peaaegu tasakaalus. Aas-



Joonis 2.

tatel 1821–1840 prevaleerib tallinnakeelne vaimulik kirjandus ning alles ajavahemikus 1841–1850 tõuseb selle kõrvale olulisele kohale ka ilmalik kirjasõna. Sellest ajavahemikust tuleb taas arvestada ka lõunaeestiliste allikatega. Samal ajal näitab diagramm, et kümnendite lõikes on tartu- ja tallinnakeelse ilmaliku ja vaimuliku kirjanduse esindatus väga ebaühtlane ja seetõttu on siin väga raske komplekteerida traditsioonilist suletud korpust, mida esindab näiteks tänapäeva eesti kirjakeele baaskorpus. Tekstide valikul eristati seal kõigepealt tekstiklassid (ajakirjandus, ilukirjandus, teaduskirjandus jne). Järgnevalt otsustati, kui palju sõnesid igast tekstiklassist peaks korpuses olema (arvestades tekstide levikut või tiraaži). Konkreetsed tekstilõigud valiti igast tekstiklassist juhuslike arvude generaatori abil.[20]

Vana kirjakeele puhul muudab niisuguse valikumeetodi mõttetuks tekstide vähesus ja asjaolu, et korpus peaks võimaldama teha järeldusi ka eri autorite keele kohta. Näiteks on XVIII sajandi ilmalikku kirjandust ainult u 350 000 sõnet kuuelt autorilt, XIX sajandi esimese poole lõunaeestikeelset vaimulikku kirjandust kõigest u 300 000 sõnet 17 autorilt, XVII sajandi esimese poole lõunaeesti keelt esindavad peamiselt Joachim Rossihniuse teosed. Selletaolise valimi tegemiseks on piisavalt õigupoolest ainult XIX sajandi esimese poole põhjaeestikeelset vaimulikku kirjandust: u 1,5 miljonit sõnet.

Võimalik oleks korpusesse võtta teatava pikkusega lõik igast teosest või igalt autorilt. Kui tahta, et korpus annaks küllaldast infot erinevate autorite kohta, peaks võetav lõik olema piisavalt pikk, nt 20 000 sõnet. See annaks korpuse mahuks u 3 miljonit sõnet, mis on pool üldkogumi sõnede arvust. On küsitav, kas tekstilõikude valimisele kuluv aeg tasub end sel juhul ära. Ühe miljoni sõnelise korpuse annaks 5000 sõnet igast tekstist, kuid 5000 sõnet on

[20] T. Hennoste, K. Muischnek, H. Potter, T. Roosmaa, Tartu Ülikooli eesti kirjakeele korpus: ülevaade tehtust ja probleemidest, lk 587–600; T. Hennoste, K. Muischnek, Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: Tartu Ülikool, 2000, lk 183–217.

autori või teksti keeletarvituse kohta järelduste tegemiseks ilmselgelt liiga vähe. Ka statistiline analüüs oleks niisuguse korpuse põhjal üsna keerukas.

Kõige otstarbekam tundub siiski koondada valikkorpusesse iga tekstiklassi olulised tekstid. Olulisuse kriteeriumideks võiks olla esmasus, nt esimene katekismus, jutlustekogu, ilukirjandusteos, perioodiline väljaanne jne; kultuurilooline tähtsus, nt esimene täispiibel või uus testament; kordustrükkide olemasolu; suur tiraaž (mida kohe ei hävitatud). Näitena olgu esitatud sellistest kriteeriumidest lähtuv esialgne, s.t ka tööjõuvõimalusi arvestav tekstivalik XVIII sajandi kohta.

XVIII sajandi korpusetekstid

1715	”Meie Jssanda Jesusse Kristusse Uus Testament”	eP	vaimulik kirj	532 lk
1732	A. Thor Helle ”Anweisung…” (10 dialoogi)	eP	ilukirj	24 lk[21]
1739	”Piibli Ramat”	eP	vaimulik kirj	1395 lk
1740	”Wiis head jutto”	eP	vaimulik kirj	68 lk
1764	”Önsa Lutterusse Katekismus”	eL	vaimulik kirj	88 lk
1766–67	A. W. Hupeli ”Lühhike öppetus…”	eP	perioodika	159 lk
1771	A. W. Hupeli ”Arsti ramat”	eP	tarbekirj	162 lk
1776	J. Chr. Quandt ”Kolm Kaunist Waggause Eenkojut”	eL	vaimulik kirj	60 lk
1779	”Juttusse-Ramat” J. B. Sczibalsky	eL	vaimulik kirj	788 lk
1779	”Jutlusse Ramat”	eP	vaimulik kirj	652 lk
1781	”Köki ja Kokka Ramat”	eP	tarbekirj	699 lk
1782	R. W. Willmanni ”Juttud ja Teggud”	eP	ilukirj	227 lk
1782	F. G. Arweliuse ”Üks Kaunis Jutto ja Öppetusse…”	eP	ilukirj	126 lk
1787	sama II jagu	eP	ilukirj	152 lk
1790	”Lühhikenne öppetus maa-rahwale”	eP	tarbekirj	32 lk
1792	A. Raudialli ”Ütte wanna Jesusse Teenre…”	eL	vaimulik kirj	40 lk
1793	G. G. Marpurgi ”Kristlik Oppetusse-Ramat”	eL	vaimulik kirj	202 lk
1796	F. D. Lenzi ”Aija-Kalender”	eL	tarbekirj	64 lk
1799	G. G. Marpurgi ”Ma-rahwa Laste-kaswatamissest”	eL	tarbekirj	16 lk

Olulisuskriteeriumi alusel valitud XVIII sajandi korpuseteksti on kokku 5486 lk, sellest suurima osa moodustavad põhjaeestiline vaimulik kirjandus (2647 lk) ning lõunaeestiline vaimulik kirjandus (1178 lk). Põhjaeestilist tarbekirjandust on valimis 893 lk ja lõunaeestilist vastavalt 80 lk. Põhjaeestikeelset ilukirjandust on 529 lk, lõunaeestilist ilukirjandust ei ole.[22] Esindatud on ka esimene perioodiline väljaanne A. W. Hupeli ”Lühhike öppetus…”.

XVIII sajandi kohta koostatud esmasest tekstivalikust hoolimata on töörühma kaugemaks eesmärgiks siiski kõiki vana kirjakeele trükitekste sisaldav korpus.

[21] Arvestatud on vaid eestikeelset teksti.

[22] XVIII sajandi kontekstis on ilukirjanduse ja vaimuliku kirjanduse eristamine mõneti tinglik. Siinses ülevaates on piiblitemaatikaga manitsevad proosapalad loetud vaimuliku kirjanduse näideteks.

Kokkuvõte

Praeguseks on selge, et infotehnoloogia hõlbustab oluliselt keelematerjali talletamist ja analüüsi ning elektrooniliste materjalikogudeta on tänapäevast süsteemset keeleuurimist võimatu ette kujutada. See aga, kui suures mahus ja missuguseid tekste peaks sisaldama vana kirjakeele korpus, sõltub ennekõike uurijate huvist ja uurimiseesmärkidest, korpuse täienemise kiirus oleneb kahtlemata ka kasutada olevast tööjõust.

Grammatikanähtuste süsteemseks uurimiseks peaks korpus koondama eesti kirjakeele ajaloos oluliste autorite tekstivalimiku. Esindatud peaksid olema tallinna- ja tartukeelsed tekstid ning eri registrid (vaimulikud tekstid, tarbetekstid, ilukirjandus, juriidilised ja ajakirjandustekstid). Vana kirjakeele korpus peaks uurijate käsutusse andma kergesti käsitsetava ja usaldusväärse andmekogu.

Tulevikus oleks ideaalne jõuda vana kirjakeele kõikse korpuseni, mis koondaks kõiki kirjalikke tekste. Et see ei ole praeguste võimaluste juures lähema kümne aasta jooksul teostatav ülesanne, tuleks kohe alustada valikkorpuse loomist, mida saab hiljem vajaduse korral laiendada.

Vana kirjakeele korpuseta ei ole meil võimalik tõusta keelenähtuste uurimisel üksiktähelepanekutest kõrgemale üldistustasandile ega analüüsida kirjakeele arengut selle järjepidevuses. Tartu Ülikooli eesti keele õppetooli vana kirjakeele uurimisrühmas on praeguseks välja töötatud vana kirjakeele korpuse koostamispõhimõtted ning loodud olemasolevatest trükiteostest andmebaas. On olemas ka väljaõppinud inimesed, kes suudavad korpusematerjali usaldusväärselt sisestada ja märgendada, ning võimalus uusi töötajaid ette valmistada – seega igati head eeldused, et kirjakeele ajaloo materjali talletada ja uues kvaliteedis mõtestada.