Eesti keele kui emakeele õppija tekstikorpus EMMA

Kadri Sõrmus, Kersti Lepajõe
Tartu Ülikool

Ülevaade. Artiklis tutvustatakse Tartu Ülikooli eesti ja üldkeeleteaduse instituudis loodavat korpust EMMA, mis koondab eesti keelt emakeelena õppivate kooliõpilaste autentseid tekste (nt riigieksamitööd, põhikooli lõpueksami tööd, uurimistööd). Korpuse eesmärk on luua tänapäevased võimalused õpilastekstide erijoonte, õpilaste tekstiloomeoskuste, keelekasutuse muutuste, emakeeleõpetuse, eksamivormi jm uurimiseks. Artiklis antakse ülevaade korpuse EMMA loomise teoreetilistest lähtekohtadest (J. Sinclairi, D. Biberi, S. Hunstoni, R. Reppeni jt põhjal), korpuse eripärast, tutvustatakse korpuse loomise protsessi ning korpuse rakendusvõimalusi. Korpuse EMMA loomise praeguses järgus digitaliseeritakse, trükitakse ja esmamärgendatakse 3000 tööst koosnev valim perioodil 1998–2014 kirjutatud riigieksamikirjandeid ning luuakse kasutajasõbralik veebikeskkond.

Võtmesõnad: keelekorpused, õpilaste keelekasutuse uurimine, keeleõpe, emakeeleõpe, testimine, hindamine, eksamiarendus

Tartu Ülikooli eesti ja üldkeeleteaduse instituudi töörühmas alates aastast 2008 kujundatav emakeeleõppija tekstikorpus EMMA[1] on keskkond, mis koondab eesti keelt emakeelena õppivate kooliõpilaste õpieesmärgil koostatud tekste. Elektrooniline tekstikogu pakub võimaluse tänapäevaseid meetodeid kasutades uurida autentsete kirjutiste kaudu õpilaste keele- ja tekstiloome iseärasusi. Korpuse

[1] EMMA loomist toetab Haridus- ja Teadusministeerium.

loomine on aja- ja töömahukas ning koosneb paljudest omavahel seotud etappidest, nagu valimi koostamine, tekstide skaneerimine, trükkimine, korpusesse sisestamine ning märgendamine. Artikkel annab ülevaate uue korpuse EMMA koostamise põhialustest, iseärasustest, märgendamis- ja otsinguvõimalustest.

Korpuse loomise lähtekohti

John Sinclairi definitsiooni kohaselt on korpus lingvistilise uurimise eesmärgil koostatud representatiivne elektrooniline tekstikogu (Sinclair 2004). Õppijakeele korpuste koostajatel on tekstide valiku üheks kriteeriumiks autorid kui keeleõppijad. Õppijate keelt sisaldavaid ingliskeelseid korpusi hakati koostama alates 1990. aastatest ehk u 30 aastat pärast üldkorpuste loomist (Nesselhauf 2005: 40). Praeguseks on erinevatel eesmärkidel loodud ja erinevaid keeli sisaldavatest õppijakeele korpustest saanud keelekorpuste maailmas terve omaette suund paljude eriilmeliste ja spetsiifiliste korpustega (vt Learner Corpus Association[2]).

Enamasti käsitatakse õppijakorpust teise keele (L2), võõrkeeleõppija tekstikorpuse või vahekeelekorpusena, millele emakeelse õppija (L1) tekstid on lisatud võrdlusalusena (Bowker, Pearson 2002: 13 jt), kuid on ka vaid emakeelt sisaldavaid korpusi (nt LOCNESS the Louvain Corpus of Native English Essays[3]). Keelekorpuste maht on väga kõikuv ulatudes väikestest 2000-sõnalistest kuni mitmete allkorpuste liitmise teel loodud miljoneid sõnu sisaldavate õppijakorpusteni. Valdavalt sisaldavad need inglise keele õppijate tekste (Pravec 2002: 90).

---

2 http://www.learnercorpusassociation.org/resources/

3 http://www.learnercorpusassociation.org/resources/corpora/locness-corpus/

Tänapäeval on uurijatel korpuse loomiseks väga soodsad võimalused. Enamasti koostatakse korpus kättesaadavatest tekstidest ning otsus, mida korpusesse lisada ja mida mitte, sõltub eelkõige korpuse eesmärgist ning sellest, mida korpuse abil uurida tahetakse (Hunston 2012: 26, Atkins jt 1991). Nt kirjaliku keele korpused võimaldavad eelkõige uurida ning arendada kirjutamis- ja lugemisoskust (Eslon, Metslang 2007: 101), tekstikatkenditest koosnevad korpused uurida grammatikat, terviktekste sisaldavad korpused lubavad analüüsida teksti struktuuri (Meyer 2009), õppijakorpused aga keele omandamise erijooni jne. Tekstide valiku põhimõtetest sõltub korpuse sisu. J. Sinclair (2004) väidab, et korpuse ülesehitusel tuleb lähtuda võimalikult vähestest selgelt eristavatest kriteeriumitest, mis piiritlevad tõhusalt keelevaldkonna, mida korpus hakkab esindama. Üldkeelekorpuse eesmärk on anda keelest üldpilt ning seetõttu peab see sisaldama võimalikult paljusid erinevaid žanre ja teemasid õiges proportsioonis. Allkorpuste tekstivaliku põhimõtted lähtuvad nt õppijast, ülesande tüübist, tekstide kogumisviisist, tekstiliigist, keelest, teemast vm. Nn uue põlvkonna korpuste uurimisparadigmad on muutumas rakenduslikumaks (vt Eslon 2014: 437).

Sõltuvalt korpuse eesmärgist võib koguda erinevat tüüpi tekste kas eraldi või teatud valimina. Näiteks sisaldab USA Michigani ülikooli kirjalik tekstikorpus MICUSP seitset tüüpi kirjalikke tekste 16 valdkonnast ning neljal erineval õppija oskustasemel (MICUSP). Spetsiifiline korpus võib sisaldada ka vaid ühe žanri, nt akadeemilise keelekasutuse tekste (Meyer 2014: 44). Kui korpus sisaldab ainult ühte tekstitüüpi (nt ainult õpilaste esseelaadseid arutlevaid kirjandeid), tuleb tulemuste tõlgendamisel arvesse võtta, et analüüsitulemused võivad iseloomustada mitte sihtgrupi teksti eripära üldiselt, vaid pigem konkreetse žanri keelekasutust (Barlow 2005: 336).

Tekstide keelt mõjutavad autori vanus, sugu, haridustase, keelekeskkond, emakeel, päritolu, sotsiaalne kuuluvus, teised õpitavad

keeled ning nende oskuse tase, varemõpitud keeled ja keele prestiižsus, autori õpistiil, motivatsioon, suhtumine (Barlow 2005: 337, Tono 2002: 1), üldine lugemus, kirjutamise eesmärk jm. Keelenäidete kogumisel võib ka pealtnäha iseenesestmõistetav tekitada küsimusi. Näiteks peavad täiskasvanute keelekasutust koguvad korpuse koostajad otsustama, mis vanusest alates lugeda inimest täiskasvanuks (Meyer 2014: 50). Samalaadsed otsused tuleb teha ka emakeele, teise keele, haridustaseme, sotsiaalse kuuluvuse jt määratluste kohta.

Kui korpusesse lisatakse tekste lähtuvalt nende loomise situatsioonist, tuleb arvestada õppijale antud ülesande, selle kirjutamiseks antud aja, eesmärgi, spontaansuse, abimaterjalide kasutamise võimaluse, hinnatavuse ning ajapiiranguga (Tono 2002: 1, Barlow 2005: 337). Õppijakorpuste koostamisel on oluline metaandmetes märkida, kas õppija kirjutis lähtub mõnest alustekstist või on teema vaba. Analüüsi käigus tuleb alustekstide mõju arvesse võtta.

Keelekorpuse tekstide üheks oluliseks tunnuseks on autentsus (Dash 2008). Õppijatekste sisaldavate korpuste puhul võib autentsus olla problemaatiline, kuna õppijad on situatsioonis, kus nad on sunnitud teksti looma (Nesselhauf 2005: 40). Eksamisituatsioonis kirjutamine on piiratud aja, teema, alustekstide, istekoha vahetamise ja arvuti kasutamise keelu jt teguritega, mis on autentsele tekstiloomeolukorrale võõrad. Seega võib väita, et tekstid on sedavõrd autentsed, kuivõrd autentne on koolitekstide kirjutamise situatsioon, kuid autentsus kirjalikus õppijakorpuses ei ole samavõrd iseenesestmõistetav kui nt suulises vestluskorpuses. Õppijakorpuse tekstid on autentsed koolikeskkonna seisukohalt.

Sageli on korpuse koostamisel üheks oluliseks tekstide valimise kriteeriumiks teksti kättesaadavus (Hunston 2012: 26) ning selle kopeerimise ja kasutamise seaduslikkus (Meyer 2014: 37). Publitseeritud vabaks kasutamiseks mõeldud tekste, mida on võimalik

veebist alla laadida, saab korpusesse lisada lõputul hulgal. Paberil käsikirjatekstide sisestamine on kulukas ja aeganõudev, litsentseeritud tekstide korpusesse lisamist piirab kasutusloa hankimise vajadus. D. Biber jt (1998: 179) on kirjeldanud õpilastekstide leidmist ja lisamist korpusesse keeruka ülesandena, kuna pole täpselt teada, mis allosi peaks õppijakorpus sisaldama, ning teisalt pole veebis ka piisaval hulgal õpilaskirjutisi, mida korpusesse kopeerida.

Sünkroonilised korpused peavad iseloomustama keele hetkeolukorda ning seetõttu sisaldama võimalikult värskeid tekste. Keelemuutused võivad toimuda suhteliselt lühikese aja jooksul. Diakroonilised korpused koondavad kindlalt määratletud varasema perioodi tekste (nt vana kirjakeele korpus). Seirekorpuse tekstid kogutakse kindlate vaheaegade järel kirjutatud tekstidest. Analüüsitavate tekstide hulk peab olema konstantne ning omavahel hästi võrreldav. Suletud korpuse puhul määratletakse periood, mida korpus peab iseloomustama, ning valitud perioodist varem või hiljem kirjutatud tekste korpusesse ei lisata. Avatud korpusesse võib uusi tekste lisada igal ajal.

Korpuse tekstide valimisega on tihedalt seotud autorite hulk ja tekstide arv. Ühe autori tekstide põhjal saab analüüsida kindla autori keelekasutust, paljude autorite tekstivalim võimaldab uurida keelt üldisemalt (Bowker, Pearson 2002: 49). Üldistava lingvistilise analüüsi tegemiseks peab korpus olema küllalt suur ja mitmekesine, et sisaldada piisaval hulgal keelenäiteid.

Kuigi enamasti soovitatakse korpustes kasutada seotud teksti, mitte üksikuid lauseid, ei ole korpuse tasakaalustamise nõuet silmas pidades alati võimalik lisada kogu materjali terviktekstidena. Fragmentide kasutamine võib olla vajalik, et hoida tasakaalus autorite, tekstiliikide, suulise ja kirjaliku teksti vm suhet ning mitte anda liigset kaalu ühele lihtsalt seetõttu, et tekst on teistest pikem. Korpuse

koostajad eelistavad üldjuhul lisada pigem suuremal hulgal lühemaid tekste rohkematelt autoritelt kui pikemaid tekste vähematelt autoritelt. (Meyer 2014: 38–39)

Kui korpus peab iseloomustama keelt üldiselt või allkeelt üldiselt, muutub tekstide valimine oluliseks, sest läbilõige peab esindama keelekasutust õiges proportsioonis (Atkins jt 1991). Suured korpused võivad lubada suuremat variatiivsust, kuid väiksem tekstivalik peab olema selle valdkonna suhtes võimalikult esinduslik. Kuidas täpselt tuleks representatiivset balansseeritud korpust luua, on vaieldav (Hunston 2012: 27–30). Suurim probleem on, et peaksime representatiivsuse saavutamiseks täpselt teadma, mis iseloomustab seda tervikut, mille suhtes korpust luues representatiivsust taotleme. (Atkins jt 1991, Hunston 2012: 30) Täielik representatiivsus ei ole reaalselt saavutatav (Muischnek 2006).

Douglas Biberi jt (1998) kirjeldatud representatiivsust taotleval õppijakeelekorpusel Corpus of Elementary Student Speech and Writing (koostaja R. Reppen) on kolm osa: lastele kirjutatud tekstid, laste poolt kirjutatud tekstid ja laste kõnekeel. Loodud korpust ei pea autorid ise kuigi esinduslikuks, sest puudub laste keelekasutusele iseloomulik mängusituatsioonis kasutatav keel ning teisedki valdkonnad on esindatud vaid väikeste mahtudega. Ometi annab korpus ettekujutuse, missugune võiks olla õppijakeele uurimiseks mõeldud representatiivse korpuse struktuur. D. Biberi jt sõnul (samas) peaks representatiivne õppijakeelekorpus sisaldama nii suulise kui ka kirjaliku suhtluse tekste. Õpilaste loodud tekstide kõrval väärib esiletoomist Biberi jt rõhuasetus sisendkeele lisamise ning uurimise tähtsusele, kuna õppija loob oma tekste alustekstidele toetudes. (Biber jt 1998: 176–178)

Õppijakorpuste seas on ka selliseid, mis ei ole loodud keele uurimise, vaid õppimise eesmärgil ning mille kasutajaks on uurija asemel keeleõppija. Kui korpuse eesmärk on pakkuda õppijale

võimalust keelenähtuste iseseisvaks uurimiseks, ei ole tekstide representatiivsus keele kui terviku suhtes oluline (Hunston 2012: 27). Õppija jaoks võib korpus olla ka selekteeritud ja süstematiseeritud keelenähtuste kogum, mida saab kasutada näidete või eeskujuna (Dash 2008).

Korpuse suurus ei ole enam ammu piiratud arvutite salvestamisvõimalustega, vaid pigem infotöötlusprogrammide kiirusega. Alati ei ole siiski otstarbekas kasutada mahukat korpust, mis pakub otsinguvasteks sadu tuhandeid sõnu. Väiksema andmehulga analüüs on kiirem, hoomatavam ning uurija jaoks kergemini tõlgendatav. Suure andmemahuga korpuste kasutusel on oluliseks muutunud filtreerimisvõimalus. Uurijal peab olema võimalik olemasolevast korpusest selekteerida välja ainult need tekstid, mis just teda konkreetselt huvitavad. (Hunston 2012: 25–26) Filtreerimise aluseks on metaandmed, millena korpuse koostaja peab lisama kõik, mis tal teada on, kuid uurija võib analüüsitulemusi esitades valida vaid need, mis on tema uuringu seisukohalt olulised.

Korpuse märgendamiseks nimetatakse teksti sisse sellise informatsiooni lisamist, mis võimaldab tekstides implitsiitsel kujul sisalduvat nähtavale tuua ja analüüsitavaks teha. J. Sinclair peab oluliseks, et uurijale jääks alati näha ka märgenditeta tekst. Ta rõhutab, et kuigi peame püüdlema täiuslikkuse poole, tuleb meeles pidada, et mingi osa andmetest jääb alati klassifitseerimata ja täielikku perfektsust ei ole võimalik saavutada. Isegi juhul, kui 100 mln sõna suurune korpus on 99% korrektne, on selles ikkagi enam kui miljon viga. (Sinclair 2004)

Peamiselt keskenduvad korpuspõhised analüüsid kas keele struktuurile või kasutamisele (Biber jt 2006: 269). Erikeelekorpused esindavad allkeelt ning on koostatud selleks, et nende põhjal nt sõnaraamatut koostada, lapse keele arengut jälgida või muud

uurimiseesmärgile vastavat analüüsida (Rizzo 2010: 3). Õppijakorpuste puhul peetakse enamasti silmas võimalust uurida keelekasutust, õppija vigu, keelelisi mõjutusi, keele muutumist, sõnavara jms.

Emakeeleõppija tekstikorpus EMMA

Eestikeelsete õpilastekstide korpuse loomise põhjuseks on vajadus õpilaste kirjutatud tekstide süstemaatilise ja tänapäevaseid uurimismeetodeid rakendada võimaldava elektroonilise andmebaasi järele. Korpuse loomise eesmärkideks on

- koguda eesti kooliõpilaste loodud autentseid tekste viisil, mis võimaldab piki- ja süvauuringuid
- analüüsida eesti kooliõpilaste keelekasutust ja tekstiloomeoskust
- uurida keelemuutustele viitavaid tendentse
- pakkuda uurimispõhist sisendit emakeeleõpetuse ja õppematerjalide arendustööks
- analüüsida eesti keele riigieksamit kui testiliiki ja testimisvormi ning teha ettepanekuid eksamiarendustööks.

Emakeeleõppija korpus EMMA on kavandatud kirjalike tekstide korpusena, kuhu on võimalik pidevalt tekste lisada. Lisamine peab toimuma esialgsete koostamispõhimõtetega samadel alustel.

Emakeeleõppija korpuses sisalduvate õpilastekstide keeleks on valdavalt eesti keel kui emakeel, st õpilane õpib eesti keelt emakeele või hariduskeelena ja sooritab eesti keele kui emakeele eksami. Selline keelemääratlus on tinglik, sest näiteks eesti keele kui emakeele lõpu- või riigieksami sooritajal ei ole vaja märkida oma emakeelt ja sellise töö kaasamisel korpuse koostajal emakeele metaandmed puuduvad. Pikemat aega eesti klassis õppinud õpilane võib otsustada, et sooritab eesti keele riigieksami kui emakeeleeksami isegi juhul, kui eesti keel pole tegelikult tema emakeel.

Õppijakeelekorpused esindavad mingit kindlat aspekti keelest ning on seega spetsiifilised. Kõik keelekogud peaksid olema representatiivsed selle valdkonna suhtes, mida nad esindavad. Emakeeleõppija korpus peaks sellisel juhul sisaldama läbilõiget kõikidest tekstidest (ka suulistest), mida õpilane erinevates situatsioonides loob. Veelgi enam, õppijakorpus peaks sisaldama ka tekste, mida õppijad on sisendkeelena kuulnud või lugenud analüüsitaval perioodil (vt Biber jt 1998).

Kogu õpilaskonna keelekasutust iseloomustava sisendkeelekorpuse koostamine on samas väga komplitseeritud ning ükskõik mil viisil seda koostada, tekib küsimus valimi usaldusväärsuse kohta. Sisendtekstide allkorpuse loomist piirab see, et puudub selge pilt tervikust, mille suhtes representatiivset sisendkeelekorpust luua. On küll võimalik välja arvutada, missuguses proportsioonis saab keskmine õpilane sisendi populaarteaduslikust, ilukirjanduslikust ja meediatekstist, vaba suhtluse tekstist, koolikeskkonnast, blogidest, siltidest, sõnaraamatutest, etikettidelt jne ning sellest proportsioonist lähtuvalt ka sisendtekste valida, kuid kas tulemus ikka iseloomustaks õpilaste tegelikku keelekasutust? Tänapäeval võib mõne eesti õpilase lugemus sarnaneda nt USA-s elava sõbra tekstivalikuga suuremal määral kui eakaaslastega Tallinnas ja Valgamaal. Võime ka esitada küsimuse, kes on keskmine õpilane ning kas tema tekstimaailm iseloomustab eakaaslaste sisendtekste? Keskmise õpilase profiili loomisel tuleb arvestada, et see on igal juhul tinglik ning ilmselt mitte täiesti tõene. Kui aga korpus sisaldab 10 või 15 aasta vanuseid õpilastekste, kas on üldse tagantjärele võimalik otsustada, mis võisid olla tookordse keskmise õpilase sisendtekstid ja nende tekstide proportsioonid? Sisendteksti piiritlematuse ja määratlematuse tõttu korpuse EMMA loomise esimeses faasis sisendtekstide kogu loomisega ei tegeleta. Emakeeleõppija tekstikorpuse EMMA koostajate tähelepanu on õpilaste kirjutatud tekstidel e väljundil.

Õppijakorpuse EMMA tekstid

Õppijakeele tekstikorpuse EMMA puhul on tegemist eesti keelt emakeelena õppivate koolilaste õppimisprotsessis kirjutatud tekstide koguga. Tekstid on esialgu jagatud nelja kategooriasse: eksami- ja tasemetööd, õpilasuurimused, esseevõistlusele saadetud õpilastekstid ja muu tekstiloome. Õppijakorpusesse kogutakse tekste kahel tasemel: gümnaasium (11. ja 12. klass) ning põhikool (8. ja 9. klass).

Eksami- ja tasemetööd. Eksami- ja tasemetöid iseloomustab selge ülesandepüstitus, piiratud ajakasutus, pikaajaline ettevalmistus ning töö tajumine õpilase tuleviku seisukohalt tähtsana. Iga õpilane on seda teksti kirjutades andnud endast parima nii sisu, kirjandi struktuuri kui ka õigekeelsust arvesse võttes. Kirjanditeema avamise oskus sõltub nii õpilase oskusest teemat interpreteerida kui ka ainevaldkonna taustateadmistest (Lepajõe 2012: 40).

Eksamikirjandite allkorpust ahvatles looma suhteliselt piiratud ligipääsuga suurepäraselt süstematiseeritud pikaajaline tekstiarhiiv Innoves, kuhu on kogutud tekste alates aastast 1997, mil gümnaasiumi lõpus kirjutatav kohustuslik eesti keele eksam muutus riiklikult hinnatavaks. Perioodil 1997–2012 on kõik gümnaasiumilõpetajad (u 7000–10 000 noort aastas) kirjutanud eksamitööna u 600-sõnalise, alates 2012. aastast, kui eksamivorm muutus, u 400-sõnalise arutleva kirjandi. Kõik eesti keele riigieksamitööd on arhiveeritud.

Korpusesse on eelmainitud eksamitöödest kavas sisestada kuus aastakäiku: 1999, 2002, 2005, 2008, 2011, 2014. Tekstide maht ühe aasta kohta on kõikse valimi sisestamiseks liiga suur, mistõttu tuleb iga esindatud aasta kohta koostada valim. Valimi koostamisel võetakse arvesse kõiki koolitüüpe, poiste ja tüdrukute suhet, kogu hindeskaalat vastavalt eksamitulemustele, valitud teemasid, erinevaid Eesti piirkondi. Valimis ei kajastu see, kas töö on või ei

ole läbinud apellatsiooni. Ühe aastakäigu maht koguhulga suhtes peaks esindusliku valimi korral olema min 300 tööd. Korpuse EMMA valim ühe aasta kohta sisaldab 500 tööd.

Riigieksamikirjandi tekstid on käsitsi kirjutatud. Selleks, et need saaksid masinloetavaks, on vaja kõik tekstid arvutisse trükkida. Teksti trükkimisel tuleb arvesse võtta, et iga töö peab jääma originaalkujuliseks, trükkija ei tohi muuta teksti autori sõnastust ega ühegi sõna kirjakuju. Ka juhul, kui autori tekst hälbib normeeritud kirjakeelest, tuleb see trükkida originaalkujul koos vigadega.

Juhuslike trükivigade tuvastamiseks on EMMA keskkonnas võimalik sisestatud teksti kohal avada õpilastöö skaneeritud koopia, mille abil saab veenduda, et sisestatud tekst ning esmamärgendus vastavad originaalile. Vajadusel on uurijal võimalik administraatori kaudu trükitekstidesse parandusi teha. Alates 2012. aastast, mil riigieksamitööde juurde kuuluvad ka alustekstid, lisatakse alusteksti fail vastava aastakäigu tööde juurde.

Riigieksamitööde kõrval säilitatakse riiklikult iga aasta kohta ka 1500 tööd sisaldav valim 9. klassi lõpul kirjutatud eesti keele riiklikest eksamitöödest. Õpilase keeleoskuse uurimise vaatepunktist on perspektiivikas skaneerida ning sisestada ka nooremate õpilaste tasemetööd. Tasakaalu säilitamise eesmärgil peaks nooremate õpilaste tekste sisestama arvuliselt rohkem, sest need on mahult lühemad (u 200 sõna). Kuna tasemetööde säilitamiseks koostatud valimid on koolipõhised, pole veel selget otsust, kas oleks õige sisestada kogu vastavate aastakäikude säilinud valim. Tasemetööde trükkimisel kehtivad samad nõuded nagu eksamitööde trükkimiselgi.

Uurimistööd. Lisaks kooliastme lõppu tähistavatele esseelaadsetele kirjanditele loovad õpilased koolisituatsioonis veel mitmesuguseid tekste. Emakeeleõppija-korpusesse on kavas koondada ka õpilasuurimuste tekste erinevatest ainevaldkondadest. Selle allkorpuse eesmärgiks on noorte kirjutatud teadusteksti analüüsimine

ning analüüsi põhjal soovituste ja ettepanekute tegemine uurimistööde keele, ülesehituse, alustekstide kasutamise jm parandamiseks. Aastast 2011 on uurimistöö kohustuslik nii põhikooli- kui ka gümnaasiumiastmes (põhikoolis on uurimistöö võrdsustatud loovtööga, kuid ka loovtöö kohta tuleb kirjutada seletuskiri või analüüs). Seega on kavas ka uurimistöö faile koguda vähemalt kahe vanuseastme noortelt. Uurimistööde failid kogutakse elektroonilisel kujul. Kui uurimistöö kirjutamiseks on kasutatud alustekste, on võimalik need töö lõppu koondatud kirjanduse loetelu järgi üles leida. Alustekste korpusesse ei lisata.

Võistlustööd. Võistlustöid iseloomustab õppija soov saata oma töö teistega konkureerima, mistõttu on eeldatavasti tegemist võimekamate või auahnemate õpilastega. Võistlustööde puhul ei ole kindel, kas õppija on teksti kirjutanud täiesti iseseisvalt või on keegi teda selle juures abistanud. Kõik samale võistlusele saadetud tööd on kirjutatud ühel teemal. Töö pikkus on määratletud minimaalse sõnade arvuga (võib erinevatel aastatel varieeruda). Võistlustööde eeliseks teiste õppijatekstide ees on nende elektrooniline kuju. Emakeeleõppija korpuse jaos on näiteks 2008. aastal kogutud esseevõistluse töid 243 719 sõna mahus. Võistlust korraldas SA Innove ning sihtgrupiks oli 15–19aastane õpilane.

2008. aasta esseevõistluse tekstid on lausestatud ja lemmatiseeritud, pealkirjad eristatud. Metaandmestik on napp: teada on vaid teema, teksti kirjutamise aasta, autorite emakeel ja eeldatav vanusevahemik. Vajadusel võib võistlustööde tekste kasutada võrdluskorpusena. 2008. a võistlustööde põhjal on juba analüüsitud tekstide keskmist lausepikkust, sagedasemaid grammatilisi ja sisusõnu (Sõrmus 2008).

Muu tekstiloome rubriik sisaldab igapäevase koolitööna loodud tekste, analüüse, arvustusi, kirjeldusi, arvamuslugusid, tutvustusi, reklaame jm. Teistest EMMA rubriikidest eristab neid

igapäevaelus loodud tekste töö väiksem panus õpilase tuleviku seisukohalt, tekstide loomise situatsioon, variatiivsus ja suurem spontaansus. Allkorpus annab võimaluse väiksemate allkogude liitmiseks ning nendes sisalduvate tekstide kasutamiseks võrdlusalusena. Muu tekstiloome rubriiki kogutakse tekste elektroonilisel kujul. Praegu sisaldab alamrubriik 2008. aastal tegevõpetajate abil kogutud õpilastekste: 61 793 sõna gümnaasiumiõpilaste ja 12 114 sõna põhikooliõpilaste tekste. Tekstide põhjal on analüüsitud õpilaste keskmist lausepikkust ning tekstide grammatilisi ja sisusõnu (Sõrmus 2008).

Korpuse loomise protsess

Korpuse loomise esimeses faasis aastatel 2013–2016 sisestatakse emakeeleõppija korpusesse EMMA eksami- ja tasemetöid. Sel perioodil on eesmärgiks korpuse kaudu uurijatele kättesaadavaks teha 6000 teksti, neist 3000 riigieksamikirjandit (u 600 sõna iga kirjand) ning 3000 tasemetööd (u 200 sõna). Tööprotsessi olulised osad on 1) valimi koostamine, 2) valimi skaneerimine, 3) skaneeritud tööde trükkimine ja korpusesse sisestamine, 4) sisestatud tööde esmamärgendamine ja teksti kontrollimine. Esimese faasi lõpul analüüsitakse korpuse sisu ja võimalusi ning tehakse vajadusel korrektiive tekstide edasise valimise osas.

Paralleelselt töödega, mis on seotud tekstiarhiivi elektroonilisele kujule viimisega, tuleb tegeleda sobiva keskkonna arendamisega. Emakeeleõppija korpus EMMA asub aadressil **https://korpused.keeleressursid.ee/emma/**. Keskkond on kavas uurijate jaoks avada 2014. aasta lõpus.

2014. aasta kevadel on skaneeritud 4000 teksti (1000 riigieksamikirjandit; 3000 põhikooli tasemetööd), sisestatud on 520 riigieksamikirjandit. Veebikeskkonda EMMA on võimalik tekstifaile

üles laadida ning esmamärgendust teha. Arendusel on filtreerimissüsteem ja tekstide otsingu- ning analüüsiosa. Tööd on alustanud 8 sisestajat, kellest suur osa on TÜ magistrandid või doktorandid.

Ligipääs EMMA korpuse keskkonda on võimalik ainult autenditud kasutajatel. Kõik kasutajad peavad allkirjastama korpuse kasutamise lepingu ning vastutama oma tegevuste eest nii keskkonnas kui ka seal esitatud informatsiooni kasutamisel. Emakeeleõppija korpuse loojad ja kasutajad järgivad andmekaitseseadust. Tööde autoritele ei viidata ning tekste analüüsitakse vaid üldistavalt.

Märgendamine ja analüüsivõimalused

Eesti keele jaoks on loodud lemmatiseerija, ühestaja ning võimalus lisada tekstidele automaatselt morfoloogilist märgendust. Nende funktsioonide sidumine korpusega EMMA on pikemaajalises plaanis võimalik, kuid esialgu neid ei kasutata. Eesti keele jaoks loodud tehnoloogilistest rakendustest kasutatakse automaatset lausestajat, mis annab võimaluse analüüsida sõnade hulka lauses, lausete hulka tekstis ja võimaldab esitada otsisõna lause kontekstis.

Korpuse EMMA jaoks on loodud spetsiaalne tarkvara, mis võimaldab tekste sisestada, metainfot lisada, tekste märgendada ja analüüsida. 2013.–16. aastal on korpuse EMMA fookus eksami- või tasemetööna kirjutatud tekstidel, mistõttu käsitletakse siinses artiklis vaid nende tekstide jaoks loodud märgendamisvõimalusi.

Sisestatavad eksami- või tasemetööde tekstid esmamärgendatakse, st eristatakse sissejuhatused, kokkuvõtted, pealkirjad, sisestatakse eksami- või tasemetööl näha olevad parandused kui õppijate jaoks potentsiaalselt keeruka ja veaohtliku keelendi tähised. Õpetajate tehtud parandusi on samal põhimõttel märgendatud ka Sloveenia kirjalikus õppijakeelekorpuses: „… õpetajate parandused kujutavad endast õpilaste keeleprobleemide analüüsi ning näitavad,

mida peavad õpetajad õpilaste tekstides problemaatiliseks. Me soovime seda väärtuslikku infot säilitada ning analüüsitavaks muuta." (Šolar)

Kuna kirjanditel on näha kahe riikliku hindaja märgitud vead, on kavas need esmamärgendamise käigus tekstis selliselt esile tuua, et oleks võimalik eristada 1. ja 2. hindaja märkusi. Vigade märgendamise aluseks on võetud eksamikomisjoni loodud vigade kategoriseerimise süsteem, mille alusel eristatakse kuut tüüpi vigu: stiili-, interpunktsiooni-, õigekirja-, faktiviga või küsitavus, sõna või lauseosa puudu, taandrida tähistamata (Eristuskiri 2013).

Lisaks riiklikult määratletud veasüsteemile on korpuses võimalik ära märkida

- vead, mis on näidatud teksti sees, kuid mida hindaja ei ole pidanud vajalikuks küljele välja tuua,
- esmamärgendaja poolt leitud vead vm tähelepanekud, mis on algses tekstis olemas, kuid mis ei ole parandajate poolt tähistatud.

Märgendamise protsess on kasutajasõbralik ja loogiline. Veatüübid on tähistatud erinevate tähtede ning värvidega. Märgendaja peab esmalt sõnale klõpsamise abil valima, kas tähistab esimese või teise hindaja tähistatud vigu (originaaltekstis eristatud punase ja rohelise kirjaga). Konkreetse hindaja veamärke sisestamiseks tuvastab märgendaja originaalteksti äärele toodud tähise järgi veatüübi, seejärel markeerib hiire abil vigase tähe, sõna vm ning klõpsab vastavale veatüübile värvilisel valikuribal lehe ülaservas. Tähistused jäävad teksti sees värviliselt näha.

Märgendamisotsuse tegemisel oli arutlusel ka õpilaste eneseparanduste märkimine teksti sisse. Praeguses korpuse loomise etapis on otsustatud eneseparandusi korpuse tekstides mitte markeerida, kuna sisestajate käsutuses on puhtandid ning vaid sealseid

parandusi analüüsides võib uurija jõuda valedele järeldustele (mustandeid ei ole säilitatud).

Soovi korral on igal uurijal võimalik ise uus(i) märgend(id) luua ning neid oma valitud tekstides rakendada. Uurijal on võimalik kasutada kõiki EMMA jaoks programmeeritud analüüsivõimalusi sõltumata teksti kategooriast. Kõik teksti kohta käivad metaandmed on uurijate jaoks kättesaadavad, kuid esitatakse analüüsitavast tekstist eraldi, et mitte kujundada eelhoiakut analüüsitulemuste tõlgendamiseks. Filtreerimisvõimalus on loodud vastava kategooria metaandmete põhjal.

Riigieksamikorpuse puhul on iga teksti juures selle kirjutamise aasta, teema või eksamivariant, teksti kirjutamise piirkond, autori sugu, koolitüüp ja tulemus. Tulemuse puhul on korpuse loojate käsutuses kolmikjaotus: kas töö kuulub 30% kõrgema, keskmise või 30% madalama punktiarvuga tööde hulka (hea, keskmine, nõrk). Metaandmetena on jäädvustatud ka teksti parandajate koodid ja alusteksti nõudva kirjutise puhul viide vastavale tekstile.

Olümpiaaditööde puhul on kirjutaja metaandmed need, mis on töö tiitellehel näha (õppeasutus, töö autori klass ja sugu, juhendaja, töö liik, valdkond ning kirjutamise aeg). Metaandmed lisatakse töö juurde selliselt, et uurijal ei ole võimalik teksti koos metaandmetega vaadata. Kasutatud kirjanduse loendi kaudu on võimalik tutvuda alustekstidena kasutatud tekstidega. Olümpiaaditööde alustekste korpusesse ei lisata.

Võistlustööde puhul on uurijal filtreerimiseks võimalik kasutada selliseid metaandmeid nagu kirjandi teema, võistluse toimumise aasta, teksti autori eeldatav vanusevahemik ning emakeel.

Muu tekstiloome puhul on võimalik metaandmeid koguda vastavalt vajadustele ning võimalustele. Metaandmetena tuleks tekstile lisada samad andmed nagu uurimistöö tekstide puhul. Olemasolevate muu tekstiloome tekstide metaandmestik ei ole ühtlane.

Analüüsi tulemused esitatakse sõltuvalt päringust arvandmetena, protsentidena või visualiseeritud kujul. Sõnaotsingu puhul kuvatakse otsitav sõna kas ilma kontekstita, 1–2sõnalises kontekstis või lauselises ümbruses.

Uurimissuunad

Kirjutamisoskus on eesti keele ja kirjanduse õpetamise, aga ka kogu kooli kontekstis erakordselt oluline oskus, mille mõju ulatub üksikisiku tasandilt rahva ja ühiskonna eduka toimimise tasandile. EMMA korpus võimaldab mitmekülgseid noorte kirjaliku keele alaseid uuringuid.

Esmalt pakub uurijatele huvi eesti keele riigieksamikirjandite põhjal koostatud allkorpus, sest see sisaldab tekste aastatest 1997–2014. Korpuses sisalduvate kirjanditekstide alusel saab analüüsida koolilõpetajate tekstiloomet: kas ja kuivõrd vastavad kirjandid tekstiliigitunnustele, millised on argumenteeriva teksti (kirjand väidetavalt seda on) loomise strateegiad, kas tegemist on sidusa tekstiga ja kui ei, siis millised vajakajäämised teksti sidususe osas ilmnevad. Selle kõrval saab analüüsida tekstide retoorikat ja interpersonaalse funktsiooni avaldumist ehk kuidas õpilased kirjutavad end tekstiosaliseks, kellega polemiseerivad, kellega suhtlevad kirjandit luues jne.

Kirjandite analüüsimine on oluline seetõttu, et sellise tekstiliigi kirjutamine eeldab õpilastelt komplekseid oskusi: lisaks žanritunnuste valdamisele on vajalik argumenteerimisoskus, võime luua tervikteksti, hallata eesti keele lausestruktuure, oskus vastavalt kehtivale õigekirjanormile kirjutada. Korpusepõhised järeldused selles vallas pakuvad vajalikku sisendit eesti keele õpetamise sisu ja vormi muutustele, õppekava ja õpivara arendusele, aga kindlasti ka õppijate keeleoskuse testimisvõimaluste avardamisele. Näiteks

2012. aastal rakendatud uue eksamitüübi lugemis- ja kirjutamisülesannete analüüs (EMMA 2014. aasta valim) võimaldab teha järeldusi eksami valiidsuse, hindamisskaalade ja kogu eksami kui testi efektiivsuse kohta ning analüüsida eksamiülesannete ja tulemuste statistilist seotust, rääkimata tulemuste soolisest eripärast.

Riigieksamikirjandid annavad hea ülevaate ka kirjakeele arengutendentsidest ja keelemuutustest, mis 15 viimase aasta jooksul paratamatult on toimunud. Noorema põlvkonna keeletaju ja selle laad vajavad analüüsimist, sest emakeeleõpetuses tuleb uutele ilmnenud tendentsidele tähelepanu pöörata, kas või suulisuse ilmingutele kirjakeeles, rääkimata morfoloogiamuutustest või teiste keelte mõjust eesti keele süntaksile. Riigieksamikirjandeid on varem uurinud Kersti Lepajõe (2012), loomulikkuse printsiipidega seotud muutusi noorte keeles aga Külli Habicht, Leelo Keevallik ja Ilona Tragel (2006: 609–625). Korpus pakuks edasisteks uuringuteks head alusmaterjali. Samas on korpuse tekstid tänuväärne materjal kirjakeele normi taustauuringuteks, sest normimuudatused peavad käsikäes käima keele sisemiste muutustega, mis noorte kirjalikus keeles ehk kõige selgemini ilmnevad (vt ka Meyer 2014: 45–46).

Sõnavara rikkus on üks parameetreid, mis iseloomustab vilunud keelekasutajat. Korpus võimaldab teha kvantitatiivseid ja kvalitatiivseid uuringuid gümnaasiumilõpetaja sõnavara rikkuse osas. 5., 7., 9. ja 11. klassi õpilaste loovkirjutiste sõnavara ja osati ka süntaksi uurimisel on huvitavate tulemusteni jõudnud Krista Kerge uurimisrühm (2014). Korpus võimaldab teha järeldusi gümnaasiumilõpetaja sõnavara rikkuse kohta, aga võimaldab analüüsida ka sõnavara individuaalsust, kui ühitada põhikooli ja gümnaasiumi eksami tulemused isikuti.

Kiiresti muutuvas maailmas on just kooli tähendus ühiskonna ühisväärtuste kujundamisel väga oluline. Õpilaste loodud tekstidest ilmnevad nende põlvkonna sotsiaalsed ja individuaalsed väärtused

ja nende muutused ajas. EMMA pakub väärtusuuringuteks vajalikku tekstikogu, mis võimaldab viimase 15 aasta arengutendentse jälgida.

Kokkuvõte

Aastani 2014 ei olnud eesti keelt emakeelena kõnelevate õpilaste keelekasutuse analüüsimiseks ühtegi selgetelt alustelt koostatud elektroonilist tekstikogu, mis oleks võimaldanud kiireid otsinguid ning tänapäevaste uurimismeetodite rakendamist. Kiiresti muutuvas maailmas on just kooli tähendus ühiskonna ühisväärtuste kujundamisel väga oluline. Seega on eesti keele kui emakeeleõppija tekstikorpuse loomine oluline samm nii üliõpilaste kui ka teiste uurijate jaoks usaldusväärse algmaterjali tagamisel ja analüüsivõimaluste mitmekesistamisel. Õpilaste loodud tekstidest ilmnevad nende põlvkonna sotsiaalsed ja individuaalsed väärtused ja nende muutused ajas. EMMA pakub väärtusuuringuteks vajalikku tekstikogu, mis võimaldab viimase 15 aasta arengutendentse jälgida.

Korpuse loomise töö on aja- ja töömahukas ning koosneb paljudest omavahel seotud etappidest, nagu valimi koostamine, tekstide digiteerimine, trükkimine, korpusesse lisamine ning märgendamine. Samuti nõuab see töö spetsiaalset veebilehte ning üsna suurt kapatsiteeti andmemahtude säilitamiseks. Emakeeleõppija tekstikorpus EMMA täidab loodetavasti temale pandud ootusi ning aitab kaasa eesti keele kui emakeeleõppija tekstide uurimisele ja uurimistulemuste kaudu ka emakeeleõpetuse ning õppematerjalide loomise kvaliteedile.

Kirjandus

Atkins, Sue; Jeremy Clear, Nicholas Ostler 1991. Corpus Design Criteria. http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf (22.05.2014).

Barlow, Michael 2005. Computer-based analyses of learner language. – Ellis, Rod, Gary Barkhuizen (eds.), Analysing Learner Language. Oxford: OUP, 335–357.

Biber, Douglas; Susan Conrad, Randi Reppen 1998. Corpus Linguistics: Investigating Language Structure and Use. (= Cambridge Approaches to Linguistics.) Cambridge: CUP.

Bowker, Lynne; Jennifer Pearson 2002. Working with Specialized Language: A Practical Guide to Using Corpora. New York, Oxon: Routledge.

Dash, Niladri Sekhar 2008. Corpus Linguistics. An Introduction. http://www.eolss.net/sample-chapters/c04/e6-91-17.pdf (28.05.2014).

Eristuskiri 2013 = Eesti keele riigieksami eristuskiri. Riigieksamite materjalid 2013. Tallinn: Innove, 2013. http://www.ekk.edu.ee/vvfiles/0/Eesti_keele_riigieksami_eristuskiri4.pdf (26.05.2014).

Eslon, Pille 2014. Eesti vahekeele korpus Estonian Interlanguage Corpus. Keel ja Kirjandus, 6, 436–451.

Eslon, Pille; Helena Metslang 2007. Õppijakeel ja eesti vahekeele korpus / Learner language and Estonian Interlanguage Corpus. – Eesti Rakenduslingvistika Ühingu aastaraamat / Estonian Papers in Applied Linguistics, 3, 99–116.

Habicht, Külli, Leelo Keevallik, Ilona Tragel 2006. Keele muutumine kasutuskontekstis / Change of language in usage context. – Keel ja Kirjandus, 8, 609–625.

Hunston, Susan 2012. Corpora in Applied Linguistics. The Corpus as Object: Design and Purpose. Cambridge: CUP, 25–37. http://www.cambridge.org/servlet/file/store7/item5633070/version1/Hunston%20Chap%202.pdf (26.05.2014)

Kerge, Krista; Anne Uusen, Halliki Põlda 2014. Teismeea loovkirjutiste sõnavara ja selle hindamine Teenage vocabulary and its assessment in creative writing. Eesti Rakenduslingvistika Ühingu aastaraamat / Estonian Papers in Applied Linguistics, 10, 157–175.

Learner Corpus Association = http://www.learnercorpusassociation.org/resources/ (20.05.2014)

Lepajõe, Kersti 2012. Kirjand kui tekstiliik. Riigieksamikirjandite tekstuaalsed, retoorilised ja diskursiivsed omadused. (= Dissertationes philologiae estonicae Universitatis Tartuensis, 31.) Tartu: Tartu Ülikooli kirjastus.

Meyer, Charles F. 2009. Planning the construction of a corpus. – English Corpus Linguistics: An Introduction. (= Studies in English Language.) Cambridge: CUP, 30–54.

MICUSP = Mishigan Corpus od Upper-Level Student Papers. http://micusp.elicorpora.info/ (26.05.2014).

Muischnek, Kadri 2006. Verbi ja noomeni püsiühendid eesti keeles. (= Dissertationes philologiae estonicae Universitatis Tartuensis, 17.) Tartu: Tartu Ülikooli kirjastus.

Nesselhauf, Nadja 2005. Collocations in a Learner Corpus. Amsterdam, Philadelphia: John Benjamins.

Pravec, Norma 2002. Survey of learner corpora. IACME Journal, 26, 81–114.

Rizzo, Camino Rea 2010. Getting on with Corpus Compilation: from Theory to Practice. English for Specific Purposes World, 9, 1 (27). http://www.esp-world.info (20.04.2014).

Sinclair, John 2004. Corpus and texts. Martin Wynne (ed.), Developing Linguistic Corpora: a Guide to Good Practice. http://www.ahds.ac.uk/creating/guides/linguistic-corpora/index.htm (26.05.2014).

Šolar = Šolar Learners' Corpus. Sloveenia. http://eng.slovenscina.eu/korpusi/solar (26.05.2014).

Sõrmus, Kadri 2008. Emakeeleõppija korpus ja veamärgendussüsteem. Magistritöö. Käsikiri TÜ eesti ja üldkeeleteaduse instituudis. Tartu: Tartu Ülikool.

Tono, Yukio 2002. Learner corpora: Design, development and applications. Japan: Meikai University. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.115.6849&rep=rep1&type=pdf (20.04.2014).

The Estonian native-speaking students' text corpus EMMA

Kadri Sõrmus, Kersti Lepajõe
Tartu University

EMMA[1], the Estonian language learners' text corpus being developed at the Institute of Estonian and General Linguistics of Tartu University, is an environment that gathers texts connected with study processes of students learning Estonian as a native language. The article gives an overview of the basis of compiling the EMMA corpus, its character, annotation, analysis and research opportunities.

The corpus texts fall into four categories: examination and level test papers, student research papers, essays sent to writing contests, and other texts. The student text corpus will include texts from two school levels: high school (grades 11 and 12) and middle school (grades 8 and 9).

During the first phase of corpus creation in 2013–2016, the focus of the Estonian native-speaking students' text corpus EMMA is on examination papers and level tests. Graduation essays of high school students have been collected in Estonia since 1997, when the compulsory Estonian language exam given at the end of high school started to be graded nationally. During the period 1997–2014, all high school graduates (approx. 7,000–10,000 12th graders per year) wrote 400–600 word argumentative essays as a national examination. In order to build the corpus, samples from 1999, 2002, 2005, 2008, 2011 and 2014 were selected, and texts were scanned, typed in, entered into the EMMA environment, and the first annotation was added, i.e. mistakes marked by the nationally selected graders.

By 2016, it is planned to enter at least 6,000 texts, including 3,000 national examination essays (approx. 600 words per essay) and 3,000 level tests (approx. 200 words), as well as making the corpus accessible to researchers through the EMMA environment.

---

1 https://korpused.keeleressursid.ee/emma/

So far, there are no electronic text corpora for analysing the language use of Estonian native-speaking students that enable quick searches and use of contemporary research methods. Therefore, creating a corpus of Estonian native-language learners' texts is an important step in providing researchers of students' papers and other researchers with trustworthy primary material and in creating more analysis opportunities. Hopefully, the language learners' text corpus EMMA will fulfil its goals, contribute to researching texts written by Estonian native-language students and, through research outcomes, contribute to the quality of native language teaching and teaching materials.

Keywords: language coprora, students' language use, language learning; L1 teaching, testing and assessment, test development