EESTI VAHEKEELE KORPUS

PILLE ESLON

1. Sissejuhatus

Õppijakeele korpustest saab ülevaate Louvain-la-Neuve'i katoliikliku ülikooli inglise korpuslingvistika keskuse kodulehelt, kust leiab mh esmateabe olemasolevate elektroonsete kogude (sh Tallinna Ülikooli eesti vahekeele korpuse)[1] kohta ja õppijakeele uurimise bibliograafia. Eelkõige on kajastatud inglise õppijakeele korpused, mille seas eristuvad keskuse korpusvõrgustik ning üle kahekümneaastase uurimistöö suunad. Juubelikonverentsil Louvainis 2011. aasta septembris sõnas keskuse juht Sylviane Granger, et kuigi on tehtud palju, seisab ees hulgaliselt keeletehnoloogilist arendustööd, mis nõuab aega ja ennekõike vahendeid.

Ühelt poolt on see protsess olnud seotud korpuste optimeerimise, standardiseerimise ja uute keeletehnoloogiliste vahendite ning rakenduste loomisega, nagu automaatne annoteerimine, märgendamine (sh veamärgendus), sõnaliigieristus, lemmatiseerimine, morfo- ja süntaksianalüüs. Korpusväliselt töötavatelt statistikapõhistelt programmidelt on üle mindud korpuspõhiste vahendite arendamisele. Muutused toimuvad sisuliselt iga 10–15 aasta järel, mistõttu ongi korpusvahendite ja tarkvara uuenemisel täheldatud viimase viiekümne aasta jooksul nelja põlvkonna vaheldumist (vt Anthony 2013: 151–157). Lisaks tehnoloogilise mõtte arengule ja avaratele rakendusvõimalustele on tõukejõuks kasvavad vajadused: soov kiiresti leida autentset lingvistilist teavet suuremahulistest, usaldusväärsetest ja tasakaalustatud tekstikogudest, et uurimistulemused oleksid adekvaatsed, et säiliks tasakaal keeletehnoloogia võimaluste, lingvistiliste rakenduste kasutajasõbralikkuse ja uurimistulemuste täpsuse vahel.

Teisalt on see protsess olnud seotud uurimisparadigmade muutumisega, teooriate testimise ja erinevate meetodite kasutamisega lingvistilises uurimistöös ning pedagoogilistes rakendustes. Konkreetne näide selle kohta on sünonüümne mõistepaar *vahekeel* ja *vahekeele korpus* ning *õppijakeel* ja *õppijakeele korpus*.[2] Uurimisainese korpuslingvistilisest käsitlusviisist (ingl

[1] Vt http://www.uclouvain.be/en-cecl-lcworld.html (7. XII 2013).

[2] „Mõisted *õppijakeel* ja *vahekeel* on kujunenud erinevates paradigmades, kuid üldjoontes tähistatakse mõlemaga keelevariante, mida õppijad sihtkeeles loovad. Termini *vahekeel* (*interlanguage*) võttis 1972. aastal kasutusele L. Selinker. Selleks lõid soodsa pinnase biheivioristlik keelekäsitlus, keelte kontrastiivanalüüs (lähte- ja sihtkeele võrdlus) ning arenev interferentsiteooria (lähtekeele negatiivne mõju sihtkeelele). [---] Kesksel kohal on keelevea (*error*) mõiste, veaanalüüs (*error analysis*) ja veataksonoomia .... Terminit *õppijakeel* (*learner language*) on kasutatud seoses teise keele / võõrkeele omandamisega (*second / foreign language aquisition*). [---] S. Grangeri määratluse kohaselt on paralleelselt õppijakeelekorpuse mõistega kasutusel terminid *vahekeelekorpus* (*interlanguage corpora*) ja *teise keele korpus* (*L2 corpora*). Tegemist on teise keele / võõrkeeleõppija loodud autentsete kirjalike tekstide või suulise kõnekeele näidete elektroonilise koguga, milles keelevead on märgendatud ja klassifitseeritud. Korpuse töötlemisel saab kasutada vastavat lingvistilist standardtarkvara,

corpus-based ja *corpus-driven*) on kujunenud omaette metodoloogiline suund. Ajas muutuvad ka korpuslingvistilised trendid: tekstikorpustelt suulise kõne korpusteni; ükskeelsetelt korpustelt mitmekeelsete, tõlke- ja multimodaalsete korpusteni; allkeeltelt diskursusanalüüsini; tõlkeabiprogrammidelt automaattõlkeni; keele õpetamiselt keele omandamise ja akadeemilise kirjutamiseni, viimastel kümnenditel keeleoskustasemete lingvistiliselt sisult tasemeoskuste automaatse testimiseni kuue- või kolmeastmelises süsteemis jm. Lingvistilises mõttes liigutakse sõnalt kollokatsioonile ja fraseologismile, vormilt tähendusele ja tekstilistele funktsioonidele, allkeelte spetsiifikalt autori keelekasutuse omapärale, kriitilisele diskursusanalüüsile, vormilt morfosüntaktilistele mustritele jm. Kogu protsessi mõte on selles, et olenemata kasutaja (lingvisti, tõlkija, keelehuvilise, õpetaja, õpilase jt) vajadustest, pakutakse korpuspõhiselt töötavaid kiireid ja kasutajasõbralikke vahendeid.

Analoogsed arengud on toimunud ka õppijakeele korpustes, mis on tekkinud kirjakeele korpustest ajaliselt hiljem. Üleminekud uuele tehnoloogilisele platvormile on tingitud muutustest uurimisparadigmas, millega kaasneb vajadus uute korpuspõhiste rakenduste järele. Tänaseks on nii mõnegi õppijakeele korpuse eelmise põlvkonna versioon konserveeritud ja üle mindud järgmise põlvkonna korpusele. Näiteks Bergeni ülikooli norra õppijakeele korpuse ASK (Norsk andrespråkskorpus)[3] Corpus Workbench'i platvorm valmis aastatel 2002–2009. Uue põlvkonna ASK-korpuse jaoks kasutatakse Corpuscle platvormi, mille lõi Paul Meurer (Uni Computing, Bergen) 2010. aastal (vt Meurer 2012). Selles keskkonnas saab kasutada korpuspõhiseid meetodeid, et uurida emakeele mõju norra keele omandamisele (nt Pepper 2012), ja testida keeleoskuse taset. Korpus sisaldab kümne erineva emakeelega norra keele kui teise keele õppijate esseesid (kokku 1700). Uurimistöö tugineb keele omandamise kognitiivsele teooriale (Slobini hüpotees „mõtlemine eelneb rääkimisele") ja keeletüpoloogiale. Võrgustuda soovitakse ennekõike nende keskustega, kus õpetatakse norra keelt või uuritakse teise keele omandamist ja õpetamist.

Teine näide on Cambridge'i ülikooli kirjastuses valminud maailma suurim inglise keele õppijakorpus The Cambridge Learner Corpus.[4] See elektroonne keelekogu on välja kasvanud Cambridge'i inglise keele korpusest (Cambridge English Corpus, varasema nimetusega Cambridge International Corpus), mille tegid leksikograafid sõnastike koostamise eesmärgil. Cambridge'i õppijakorpus on valminud koostöös Cambridge English'iga, sisaldab 200 000 eksamisooritust, mille on kirjutanud 148 erineva emakeelega inglise keele õppijad. Eksamitekstide analüüsiks kasutatakse kirjastuses välja töötatud veakodeerimise süsteemi, mis tunneb ära 77 vealiiki, oskab neid parandada, leida vahelejätte ja testida keeleoskust. Seetõttu saab korpust kasutada inglise keele tasemeoskuste ning igale tasemele omase sõnavara ja grammatika uurimiseks, tasemete lingvistilise sisu täpsustamiseks üleminekul alamalt astmelt ülemale.

Ka Tallinna Ülikooli eesti vahekeele korpuse (EVKK)[5] arendamisel on aeg üle minna uuele platvormile, mis analoogselt ASK-i ja Cambridge'i õppija-

---

korpusuuringutele tuginevad oma töös teise keele / võõrkeele õpetamise spetsialistid ....", vt Eslon 2007: 87–88.

[3] Vt http://www.uib.no/fg/askeladden (7. XII 2013).

[4] Vt http://www.cambridge.org/ee/elt/catalogue/subject/custom/item3646603/Cambridge-English-Corpus-Cambridge-Learner-Corpus/?site_locale=et_EE (7. XII 2013).

[5] Vt http://evkk.tlu.ee/ (12. I 2014).

korpusega võimaldaks kasutada avaramalt korpuspõhist analüüsi, määrata eesti keele kui teise keele tasemeoskuste lingvistiline sisu ja testida kirjutamisoskust. EVKK-sse on kavas luua e-õppe platvorm, kus nt iseõppijad ja õpetajad koos õpilastega saaksid kontrollida enda ja teiste kirjutatud tekstide keelelist korrektsust, saada tagasisidet sõnavara ja grammatika kohta, arendada oma kirjalikku väljendusoskust harjutuste ja enesekontrollitestide toel. See tähendaks EVKK funktsionaalsuse avardumist: vahekeele (õppijakeele) korpus saab õppijakorpuseks (analoogia Cambridge'i õppijakorpusega), laieneb korpuse sihtgrupp, muutuvad kasutajate vajadused, tekivad uued võimalused korpust reaalselt rakendada. Näiteks vajavad eesti õppijakeele ja keele omandamisprotsessi uurijad korpuspõhiseid analüüsivahendeid, mis aitavad leida vajalikku lingvistilist informatsiooni ja seda (aga ka korpuskeskkonda imporditud materjali) statistiliselt töödelda. Õpetajat ja õpilast see osa ei huvita – neile tuleb pakkuda keeleanalüüsi põhjal saadud tulemusi: vormide või sõnade kooskasutuse reegleid, näiteid ja harjutusi, nagu soome keele e-sõnastik ConLexis[6] või äsja ilmunud „Eesti keele põhisõnavara sõnastik" (Kallas jt 2014), mille veebiversioon on lubatud avada 2014. aasta sügisel.

Käesolevas artiklis tutvustatakse EVKK esimest versiooni, õppijakeele automaatse analüüsi võimalusi, korpuspõhiste meetodite rakendamist ning antakse lühiülevaade EVKK põhjal tehtud uurimistööst.

2. Eesti vahekeele korpuse tutvustus

EVKK on eesti keele kui riigikeele (teise keele) ja võõrkeele õppijate kirjalike tekstide kogu. Praegune, esimese põlvkonna versioon on valminud Tallinna Ülikooli filoloogide, haridustehnoloogide ja informaatikute koostööna. Korpust saab kasutada 1) empiirilist ja rakenduslikku laadi uurimistöös (nt eesti keele morfosüntaktilised kasutusmustrid ja leksikaalne varieerumine, õppijakeele morfosüntaktiline keerukus ja sõnavara rikkus, arengud eesti keele süsteemis, kasutusgrammatika, eesti keele omandamine, keeleoskuse järkjärguline arenemine, Euroopa Nõukogu keeleoskustasemed); 2) tulevaste õpetajate ja lingvistide koolitamisel (nt veaanalüüs, võtmesõna analüüs, sõna- ja vormisagedus, klasteranalüüs, kontrastiivanalüüs); 3) tegevõpetajate täiendõppes (nt korpuste ja sõnastike kasutamine keeleõppes, sh EVKK ja Keeleveeb[7]). Korpuse funktsioone saavad kasutada kõik, eriõigused on registreeritud kasutajatel, andmehalduril ja programmeerijal. Kuna EVKK on monitorkorpus, siis lisatakse sellesse pidevalt uusi tekste. Hetkel on korpuses 12 042 teksti, üldmaht 3 308 653 sõnet, teksti keskmine pikkus 274 sõnet (vt tabelit 1).

2.1. EVKK struktuur: alamkorpused

EVKK koosneb tuum- ja alamkorpustest. Eristatakse K2-K1 eesti keele ja K3-K1 vene keele alamkorpusi (vt tabel 1). K2 eesti keel on õppijakeel, mille tekstid on koondatud tuumkorpusse. Informandid on valdavalt vene emakeelega eesti keele kui riigikeele õppijad. Nende kirjalikud tööd on kogutud kahest regioonist, kus elab kõige rohkem vene emakeele taustaga inimesi

[6] Vt http://wiki.virtues.fi/conlexis/aika?highlight=(KategoriaSana%29) (23. IV 2014).
[7] http://www.keeleveeb.ee (20. V 2014).

Tabel 1

Eesti vahekeele korpuse alamkorpused

Alamkorpus	Tekstide arv	Sõnede arv	Teksti keskmine pikkus
K2 tuumkorpus	3207	813 814	254
K2 riiklikud eksamitööd	8092	2 080 368	257
K2 olümpiaadi tööd	63	58 614	930
K2 akadeemiline eesti keel	26	38 918	1497
K1 akadeemiline eesti keel	4	3339	835
K1 vene keel	371	210 003	566
K3 vene keel	279	103 597	371
Kokku	12 042	3 308 653	274

Eestis: Ida-Virumaalt (Narva, Kohtla-Järve, Jõhvi, Sillamäe) ja Harjumaalt (Tallinn).[8] Lisaks sisaldab tuumkorpus ka soome, inglise, saksa, ungari, leedu jm emakeelega õppijate töid, kelle jaoks eesti keel on võõrkeel. Teine K2 eesti keele alamkorpus koosneb riiklikest eksamitöödest, mis on talletatud Innove arhiivis (end Riiklik Eksami- ja Kvalifikatsioonikeskus, REKK). Digiteeritud on riigieksamite, eesti keele taseme- ja kodakondsuse eksami kirjalikud tööd, põhikooli ja gümnaasiumi õpilaste eksamitööd. Kolmas K2 eesti keele alamkorpus sisaldab iga-aastase üleriigilise eesti keele kui teise keele olümpiaadi töid, neljandas on akadeemilise õppijakeele tekstid üliõpilaste lühematest või pikematest teadustöödest.

K1 akadeemilise eesti keele alamkorpus kuulub Tallinna Ülikooli eesti teaduskeele keskusele, koosneb emakeelekõnelejate teadusartiklitest, magistri- ja doktoritöödest, mis pole seni olnud elektroonselt kättesaadavad, kuid on praeguseks digiteeritud ja ootavad alamkorpusse sisestamist.[9]

Referentskorpustena on K1 ja K3 vene keele alamkorpused selleks, et võrrelda emakeele ja teise (kolmanda) keele omandamisprotsessi sarnasust ning erijooni. Esimene sisaldab vene emakeelega gümnaasiumiõpilaste esseesid (kogutud Narvast) ja teine Innove arhiivi vene keele riigieksami kirjandeid, mille on kirjutanud eesti emakeelega vene keele õppijad, kelle teine keel on inglise ja kolmas vene keel.

EVKK koostamiseks valitud andmekogumismeetodid ei erine oluliselt teistes õppijakeele korpustes (nt tšehhi CzeSL[10] ja soome ICLFI[11]) kasutusel olevatest. Tekstide kogumine on piiratud regionaalselt, õppija ning õpetaja täidavad koos ankeedi, millest saab metateavet nii õppija kui ka teksti kohta: päritolumaa, sotsiaalne taust, vanus, sugu, emakeel, kodukeel, haridus, klassiruumis või kodutööna kirjutatud tekst, eeldatav keeleoskustase kolmeastmelises süsteemis (A-B-C) või kolme eksperdi evalveerituna ja vastavuses Euroopa Nõukogu kuueastmelise süsteemiga (A1-A2-B1-B2-C1-C2). Innove arhiivi

[8] Eesti keelesituatsiooni ja vene dominandiga Eestis läbi aegade, sh XXI sajandi alguses saab asjahuviline tutvuda nt Mart Rannuti (2008) artikli vahendusel.

[9] Teaduskeele keskust juhatab Peep Nemvalts, digiteerimist on toetanud riikliku programmi „Eesti keel ja kultuurimälu (2009–2013)" projekt „Teaduskeelekeskus (2009–2013)" ja TLÜ uuringufondi projekt „TLÜ Teaduskeele keskuse rajamine (2008–2015)".

[10] Vt http://utkl.ff.cuni.cz/learncorp (28. XII 2013).

[11] Vt http://www.oulu.fi/suomitoisenakielena/node/16078 (16. XII 2013).

digiteeritud tekstide metateabe kättesaadavus on sätestatud REKK-i ja TLÜ eesti keele ja kultuuri instituudi koostöölepinguga. EVKK käsikirjaliste tekstide autori isikuandmed on kaitstud kahepoolselt: õppija allkirjastab loa oma tekste uurimiseesmärgil kasutada, korpuse haldajad tagavad isikuandmete konfidentsiaalsuse. Täidetud ankeedid indekseeritakse ja hoitakse eraldi failidena. Erinevalt tšehhi õppijakorpusest hävitatakse EVKK käsikirjalised originaalid pärast tekstide sisestamist, sest kirjutamisprotsessi saab edukalt uurida eksperimentaalselt (nt ScriptLog programmi abil või silma- ja käeliigutusi ning närviimpulsse fikseerides). Korpuse tekstid on talletatud tekstiarhiivina. Tekstiliikidest on korpuses eelistatud loomingulist laadi esseed, riigieksamite kirjandid ja riiklike tasemetestide kirjutamisoskuse osa. Korpuse annoteerimise, lingvistilise märgendamise ning veamärgendusega seotud lahendused tuginevad multidimensionaalsuse põhimõttele; kasutajasõbralikkuse või uurimistulemuste täpsuse printsiibi (vt ka Kopotev, Mustajoki 2003) rakendamisel on lähtutud reaalsetest võimalustest ja uurimistöö vajadustest.

2.2. Mitmetasandiline statistika

EVKK statistika annab teavet alamkorpuste, kahe-kolme alamkorpuse või kogu korpuse ulatuses. Valiku teeb kasutaja vastavalt eesmärkidele. Statistika on mitmetasandiline, loendab ühelt poolt lingvistilist teavet (tekst, laused, sõnavormid, vealiigid) ja teisalt metateavet õppija ning teksti kohta. Teksti andmed on keel, tekstiliik, lausete ja sõnede arv. Eraldi peetakse tekstiliikide arvestust (essee, isiklik või ametikiri, vastus küsimusele jm). Veamärgendusega tekstides loeb statistika kokku ühesugused vealiigid (nt modaalverbide kasutamine, paronüümia, häälduspärane kirjaviis, rektsioon, ühildumine, sõnajärg, interpunktsioon, tekstiloome). Kogu lingvistiline teave on ühendatud õppija metateabega. Näiteks tuumkorpuses on tekst Eesti Vabariigi taasiseseisvumise järgse esimese presidendi Lennart Meri kuvandist Soome ajakirjanduses. Selle essee on kodutööna kirjutanud Soomes elav eesti keelt võõrkeelena õppiv kuni 26-aastane keskharidusega naine, kelle emakeel ning kodukeel on soome keel ja kes valdab eesti keelt C1-tasemel. Tekst sisaldab 2122 sõnet, koosneb 201 lausest ega sisalda veamärgendust (statistika näitab kokku 0 viga ja 0 erinevat vealiiki).

EVKK statistika loendab sõnavormide esinemust alamkorpustes või korpuses tervikuna. Sõnasagedust saab vaadata tähestikulises järjestuses (valida ladina tähestiku ja kirillitsa vahel) ning üldises esinemissageduse järjekorras (vt tabelit 2). Näiteks tuumkorpuses on kõige sagedam sidesõna *ja* (994 korda) ning *olema*-verbi indikatiivi preesensi ainsuse 3. pöörde vorm *on* (985 korda), mis keelekasutusele üldiselt omane: vrd eesti (esimene *ja* 210 439 korda, teine *on* 203 036 korda)[12] ja soome kirjakeele sõnavormide sagedusloendeid (esimene *ja* 1 232 623 korda, teine *on* 818 446 korda)[13] jne.

Tähestikulises sagedusloendis saab reastada kõik ühe sõna tekstikasutused ning leida laadivahelduseta sõnade kasutusparadigma. Laadivahelduslikke sõnu tuleb aga käsitsi lemmatiseerida, mis pole kuigi kasutajasõbralik

[12] Vt http://www.lexiteria.com/word_frequency/estonian_word_frequency_list.html (8. XII 2013).

[13] Vt http://www.lexiteria.com/word_frequency/finnish_word_frequency_list.html (8. XII 2013).

Tabel 2

EVKK tuumkorpuse sõnavormide tähestikuline ja üldine sagedusloend

Tähestikuline sagedusloend		**Üldine sagedusloend**	
Sõne	Sagedus	Sõne	Sagedus
aga	217	*ja*	994
ainult	52	*on*	985
arvan	42	*et*	371
alati	39	*oli*	352

(nt sõna *naaber*: **naaberis*, *naabri*, *naabrid*, **naabride*, *naabriga*, *naabrilt*, **naabris*, *naabrite*). Korrektsete vormide lemmad genereeritakse automaatselt, nt *arvama* < *arvasin, arvan, arvas, arvab, arvatakse, arvasime, arvama, arvate, arvake, arvame*. Vajalike tekstinäidete leidmiseks tuleb klõpsata konkreetsel vormil, misjärel kuvatakse dokumentide loend, kus vorm esineb. Avades dokumendi, näeb kasutaja vormi punasega markeerituna ning võib vajaliku pikkusega tekstinäited või terviktekstid kopeerida. Samad võimalused on ka pöördsõnastikul.

Kasutajale on avatud korpuse pöördsõnastik ja silbitaja. Pöördsõnastikus on sõnad reastatud tähestikuliselt sõnalõpu alusel (nt *a*-, *b*-, *c*-, *d*-lõpulised sõnad jne). Sõnu ja vorme saab rühmitada lõpusilbi järgi ning reastada sageduse alusel. Näiteks *-si* all kuvatakse kõikide *si*-lõpuliste sõnavormide sagedus: *si* 5, *asi* 744, *baasi* 6, **andebaasi* 1, *andmebaasi* 4, *hulgibaasi* 1, *peaasi* 8, *faasi* 2, *arengufaasi* 1, *gaasi* 35, *heitgaasi* 2 jne. Sama loendi võib kuvada tähestikulises järjekorras (**aadresi* 6, *aadressi* 13, **aadrissi* 1, *aafriklasi* 1, *abilisi* 1 jne) või sageduse alusel (*tagasi* 4332, *inimesi* 1576, *edasi* 1293, *asi* 744, *võimalusi* 666 jne). Leida saab ühesuguse algus-, kesk- ja lõpusilbiga sõnade ning vormide tekstisageduse korpuses või alamkorpuste kaupa. Teksti analüüsiaknasse võib kasutaja importida mis tahes tekste ning rakendada samu funktsioone.

2.3. Mitmetasandiline annoteerimine

Tekstide sisestamise esimene samm on metaandmete fikseerimine: kirjutise pealkiri ja/või tööjuhis (kui on olemas), teave õppija kohta (elukoht, sotsiaalne taust, vanus, sugu, emakeel, kodune keel, keeleoskuse tase ja abivahendite kasutamine), teksti keel (eesti, vene) ning liik (essee, isiklik kiri jm). Teise sammuna märgitakse alamkorpus, millesse uus tekst lisatakse, kolmanda sammuna kopeeritakse digiteeritud või e-posti vahendusel saadud tekst sisestusaknasse ja salvestatakse. Dokumendivaatesse lisanduvad automaatselt arvandmed lausete ja sõnede kohta. Kui on vaja parandada teksti sisestusapse või kontrollida kirjapilti, siis saab kasutada nuppu *Muuda* ja vajalikud täpsustused salvestada. Sisestatud tekstid jäävad valitud alamkorpuse dokumentide loendisse, statistika fikseerib alamkorpuse dokumentide hulga ja tekstide sisestamise aja. Kõiki dokumente saab sorteerida-filtreerida kuupäeva (uuemad eespool) ja veamärgenduse (märgendatud, märgendamata) järgi. Dokumentide loendist võib üle minna nii veamärgenduse kui ka teksti muutmise režiimi.

Igal tekstil on korpuses kolm vaadet: 1) Wordi dokument, milles on salvestatud tekst koos pealkirja ja tööjuhisega; 2) *unicode*-formaadis salvestatud *txt*-fail, kus pealkirjad ja tööjuhised puuduvad; 3) süntaktiliselt analüüsitud tekst. Kõiki variante saab eraldi avada ning eksportida. Dokumentide ja metaandmete esitamiseks on kasutatud *xml*-formaadi *xhtml*-versiooni, märgendite hierarhias on tarvitusel *xpath*-keel.

2.4. Mitmetasandiline lingvistiline veataksonoomia

Eesti õppijakeele veamärgenduse aluseks on mitmetasandiline lingvistiline veataksonoomia, mis rajaneb keeleüksuste hierarhial grafeem – morfeem – sõna – sõnaühend – lause – tekst kolmes lingvistilises aspektis (semantika, grammatika, pragmaatika), vt Eslon 2007: 104–108; Eslon, Metslang 2007: 106–112. Nende tunnuste alusel moodustub 18 veaklassi, mis sisaldavad konkreetseid vealiike ja alamliike – kokku 173. Korpuse veamärgendusaknas avaneb vealiigituse esmatasand: leksikaalsed, leksikaalgrammatilised, morfofonoloogilised, morfoloogilised, morfosüntaktilised, süntaktilised ja kommunikatiivsed vead.

Kui avada üks seitsmest esmatasandist, siis rullub astmeliselt lahti alamliikide hierarhia. Märkides konkreetse vealiigi ja/või alamliigi ning salvestades valiku, lisanduvad teksti nurksulud < >, mis markeerivad vea. Viies kursori vealiigi nimetusele, tuleb viga tekstis esile punaste nurksulgude vahel. Veamärgendus kajastub korpuse statistikas, mis loendab automaatselt erinevaid ja korduvaid vealiike ning arvutab vigade üldarvu. Tuumkorpuse tekstide praegune veamärgendus on tehtud käsitsi. Seisuga jaanuar 2014 oli tuumkorpuse üldmaht 3207 teksti, 813 814 sõnet, neist märgendatud 525 993 sõnet, vigu leitud 59 762 (11,4 %). Märgendatud vigade kohta saab teha päringuid vealiikide kaupa, mis toob esile konkordantsid (+/- üks lause). Kui vea interpreteerimiseks on vaja avarama konteksti tuge, siis saab avada tervikteksti. Tekstide veamärgendust on võimalik tagantjärele korrigeerida.

EVKK tuumkorpuse tekstide sagedamate vealiikide esikümne moodustavad sõnajärg (1358 dokumenti), interpunktsioon (1343), verbirektsioon (1239), häälduspärane kirjaviis (1105), väljendustava vastu eksimine (928), mitteafiksaalne sõnatuletus (786), noomeni ainsuse ja mitmuse vormide moodustamine ja kasutamine (740), sõnaloome ja isikupärane sõnakasutus (735), ülearune sõna lauses (694), paronüümi kasutamine (674) jne. Sageli esinevad ka põhitähenduse vead ja sõnade semantiline sobimatus; internatsionalismide mugandamise, põhikäänete moodustamise (nimetav, omastav, osastav, lühike sisseütlev), sidesõnade ja *da*-infinitiivi kasutamise, noomenirektsiooni, põhisõnaga arvus ja käändes ühildumise, aluse ja öeldise koordinatsiooni vead; lauseliikmete ärajätt, teema ja esituslaadiga seotud loogikavead. Mõnda vealiiki pole üldse märgendatud, nt onomatopoeetiliste sõnade, orientatsiooni ja seisundi afiksaaladverbide, proadverbide (seisund, otstarve ja mööndus), liittarindi ja soovlause kasutamine.

2.5. Mitmetasandiline otsing

Mitmetasandilise otsinguga kasutajaliides võimaldab kombineerida metateavet lingvistiliste andmetega, nt sõnade, sõnavormide ja vealiikidega, sortee

ühe või mitme alamkorpuse tekste kolme keeleoskustaseme (A-, B- ja C-taseme) või Euroopa Nõukogu kehtestatud kuue keeleoskustaseme süsteemis (A1-A2, B1-B2, C1-C2). Tuumkorpuse tekstide kolmeastmelise keeleoskustaseme määramisel pole kasutatud kvalifitseeritud hindajaid: enamasti on aluseks õpetajalt saadud metainfo või teksti sisestaja hinnang. Seetõttu on võetud ette tuumkorpuse tekstide ümberhindamine kuueastmelises süsteemis, mida teevad kolm kvalifitseeritud eksperti (tavapraktikas lubatud miinimum). Paari aasta jooksul on eksperdid evalveerinud ~ 1000 tuumkorpuse teksti.[14] Ümbervaatamisele ei lähe Innove arhiivi alg-, kesk- ja kõrgtaseme tekstid ning kuueastmelise süsteemi alusel hinnatud tekstide keeleoskustase, kuna nende taseme on juba määranud kõrgelt kvalifitseeritud hindajad, kelle tööd on omakorda kontrollitud.[15] Keeleoskustasemete eksperthinnang on aega ja ressurssi nõudev, kuid hädavajalik eeltöö keele omandamisprotsessi uurimiseks ja keeleoskustasemete lingvistilise sisu esiletoomiseks.

Korpuse sõna- ja vormiotsing võimaldab analüüsida kindlate vormide esinemissagedust, leida seoseid sõnavara ning vormikasutuse vahel jm. Näiteks laadivahelduseta sõna *maja* konkordantsidest tuumkorpuses tuleb esile grammatiliste käänete vormihomonüümia ning semantiliste käänete ainsuse-mitmuse vormid ehk kasutusparadigma (nt *maja-dega*, *maja-d*, *maja-s*) ja tuletised (*puu+maja-s*). Päringut võib laiendada korpusele tervikuna või valida mõne(d) kindla(d) alamkorpuse(d). Ebamugavused on seotud laadivaheldusliike sõnade tüvemuutustega. Seetõttu peab tegema mitu järjestikust päringut, mis pole kuigi kasutajasõbralik (vt ptk 2.2). Päring muutub lihtsamaks ja tulemused täpsemaks, kui korpuses hakkab tööle Kairit Sirtsi eesti õppijakeele lemmatiseerija prototüüp (vt Sirts 2012). Seni on eesti õppijakeele vormikasutuse uurimustes kasutatud muid meetodeid ja automaatanalüüsi vahendeid (vt Eslon, Matsak 2009; Eslon 2010a; 2010b; Eslon, Õim 2010; Lõo 2012; Vajjala, Lõo 2013). Need võimalused on alati olemas, kuna allkorpus(t)e tekste saab kasutajaliidese abil vabalt eksportida ja analüüsida. Tekstivalikut annab piirata ühe või mitme alamkorpuse, tekstiliigi ja keeleoskustasemega, eraldi võib analüüsida tekstide ja pealkirjade-tööjuhiste sõnastust või kasutada kõiki nimetatud võimalusi samaaegselt.

Vealiikide otsing lubab leida normivastase keelekasutuse näiteid veamärgendusega tekstidest ja siduda otsingu vajaliku metateabega. Tulemus väljastatakse konkordantsiridadena, kus viga on markeeritud punasega. Teksti nimetusele klõpsates avaneb vigade vaade terviktekstis.

2.6. Morfo- ja süntaksianalüüs, n-grammid

EVKK korpuskeskkonnas saab teha nii teksti morfoloogilist kui ka süntaktilist analüüsi. Automaatse morfoanalüüsi vahenditest töötab rahvusvahelisele standardile ja universaalsetele märgenditele[16] tuginev programm TreeTagger, mis rakendub ühtmoodi erinevatele keeltele (saksa, inglise, prantsuse, itaalia, hollandi, hispaania, bulgaaria, vene, portugali, galeegi, suahiili, slovaki,

[14] Õppijakeele tekstide evalveerimist on toetanud riikliku programmi „Eesti keel ja kultuurimälu (2009–2013)" projekt „Eesti õppijakeele tekstide hindamine Euroopa Nõukogu keeleoskustasemete alusel (2010–2013)".

[15] Hindajate hindamisest ja sellega seotud probleemidest vt Pajupuu 2007.

[16] Vt http://courses.washington.edu/hypertxt/csar-v02/penntable.html (8. XII 2013).

ladina, eesti)[17]. EVKK-s saab sellega analüüsida eesti-, soome- ja venekeelseid tekste, mis sisalduvad korpuses või on imporditud. Programmi võib soovitada erinevateks eesmärkideks: nii keelte tüpoloogilis-kõrvutavate uurimusteks, kontrastiivanalüüsiks kui ka ühe keele erinevate kasutusvariantide võrdlemiseks. Analüsaator käivitub, kui sisestada tekstid analüüsiaknasse ja valida teksti keel: eesti, soome või vene. Väljundiks on morfoloogiliselt analüüsitud tekst, mida saab eksportida, konverteerida tabeliks ja kopeerida nt Excelisse. TreeTaggeri morfoanalüüsi näide:

Minu kõige parem sõber on Marjo.

Minu	P.sg.gen	mina+0
kõige	D	kõige+0
parem	A.comp.sg.nom	parem+0
sõber	S.com.sg.nom	sõber+0
on	V.main.indic.pres.ps3.sg.ps.af	ole+0
Marjo	S.prop.sg.nom	Marjo+0
.	Z.Fst	.

See analüüsimeetod kiirendab sõnaliikide, vormide, sõna- ja vormihomonüümia ning muude tundmatuks jäänud juhtumite käsitsi ühestamist (enamasti on tegu vormimoodustus- ja lausestusvigadega, väljajättude või isiku- ja kohanimedega). Muuhulgas saab sel viisil leida seoseid sõnaliikide ja leksika vahel, kirjeldada sõnaliikide semantilist liigendust, määrata vormi- ja sõnasagedust, koostada (alam)korpuse sõnastik, leida tekstikasutuse morfoloogilised paradigmad ehk kasutusparadigmad jm. TreeTaggeri eelis seisneb selles, et programm võimaldab uurida erinevate keelte morfoloogiat samadel alustel.

Automaatse süntaksianalüüsi tegemiseks tuleb valitud tekstid sisestada aknasse ja klõpsata *uuri*. Näitelausele annab parser järgmise väljundi:

```
“<s>”

“<Minu>”
        “mina” L0 P pers ps1 sg gen cap @NN> #1->4
“<kõige>”
        “kõige” L0 D cap @ADVL #2->5
“<parem>”
        “parem” L0 A comp sg nom cap @AN> #3->4
“<sõber>”
        “sõber” L0 S com sg nom cap @SUBJ #4->5
“<on>”
        “ole” L0 V main indic pres ps3 sg ps af cap <FinV> <Intr> @FMV #5->5
“<Marjo>”
        “Marjo” L0 S prop sg nom cap @PRD #6->5
“<.>”
        “.” Z Fst #7->7
“</s>”
```

[17] Vt http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger (8. XII 2013).

Süntaksianalüsaator on integreeritud n-grammide leidmise programmiga Klastrileidja, mille on loonud Sander Ots, eeskujuks WordSmith Tools, kfNgram ja Xaira (vt Ots 2012: 6, 16), sisendiks Kaili Müürisepa parseri EstCG 1.0 väljund. Klastrileidja töötab andmekaeve põhimõttel, otsib tekstidest samalaadse morfosüntaktilise märgendusega järjendeid ja fikseerib nende esinemissageduse. Järjendite pikkust saab valida, kuid levinumad on bi- ja trigrammid, kuna erinevad uurimused on näidanud, et olenemata keelest moodustavad kahest-kolmest elemendist koosnevad sagedad lingvistilised terviküksused keskmiselt pool teksti mahust (vt Conklin, Schmitt 2008: 72–77). Vastavalt vajadusele võib otsingut piirata morfoloogiliste, süntaktiliste või morfosüntaktiliste märgendite järjenditega, arvestada saab ka (osa)lause piiri. Senistes eesti ilukirjanduskeele ja õppijakeele (võrdlevates) uurimustes on kasutatud trigramme (vt Eslon 2013; 2014; Trainis 2013; Trainis, Allkivi 2014), mille tekstisagedus on kaks ja enam korda. Samalaadse vormistruktuuriga trigrammide kogumit võib nimetada lingvistiliseks klastriks, meetodit lingvistiliseks klasteranalüüsiks.

Klastrileidja lisafunktsiooniks on käsitsi ühestamise võimalus, et morfosüntaktiline analüüs ja n-grammide leidmine oleks võimalikult täpne. Klastrileidja markeerib automaatselt mitmese analüüsi ja valib mitmesustest esimese märgenduse. Kui see osutub kontrollimisel õigeks, siis pole muuta midagi, muudel juhtudel saab olemasoleva märgenduse käsitsi parandada. Õppijakeeles on parseri märgenduse käsitsi korrigeerimine mõnikord üsna mahukas töö, kuid üldiselt pole korpuses niisuguseid raskestimõistetavaid tekste kuigi palju. Teisalt võib märgenduse parandamisest ka loobuda, sest keelekasutusmustrid toovad esile korduvad veakooslused, üksikjuhtumeid saab aga käsitleda individuaalse nähtusena seni, kuni nende kohta korpuse suurenedes uut teavet ei lisandu. Sama kehtib ka morfoanalüsaatori ehk TreeTaggeri väljundi kohta, mida samuti saab käsitsi korrigeerida.

Klastrileidja abil saadud n-grammid toovad esile keelekasutusmustreid, mis on omased eesti keele erinevatele kasutusvariantidele (nt kirjakeel *vs.* õppijakeel, murdekeel *vs.* tänapäeva kirjakeel, lapsekeel *vs.* hoidjakeel), allkeeltele (nt ajakirjanduskeel *vs.* akadeemiline keelekasutus), keeleoskustasemetele (A2-B1-B2-C1), autoritele (nt Hvostov, Kärmas või Tammsaare, Tuglas), diskursuse tüüpidele (nt poliitiline diskursus ja võim, linnadiskursus erinevate kirjanike loomingus) jne. Seega on n-grammidel avar rakendus filoloogilistes uurimistöödes, k.a tõlkevariantide võrdlemisel (nt erinevate perioodide piibli tõlked). Hetkel klastrileidjat optimeeritakse ja seotakse EVKK-süsteemiga. Seejärel saab otsida erineva pikkusega keelekasutusmustreid nii korpusest kui ka imporditud tekstidest. See võimaldab lingvistidel, keeleomandamise ja diskursuseuurijatel võrrelda eesti õppija- ja kirjakeelt samadel alustel. Õpikute autorid, õpetajad ja õpilased saavad oma käsutusse kiire mooduse leidmaks konkreetsest tekstist just selle keeleainese (keelestruktuurid, vormid ja sõnavara), millele tekstiloome tugineb. Nt õpikute koostamisel võimaldab see järgida õppeteksti, temaatilise sõnavara ja grammatika sidusust ning tekste teadlikult valida.

3. Uurimistöö

EVKK-ga seonduva uurimistöö põhisuund on eesti kirjakeele ja õppijakeele võrdlev korpusanalüüs, valdkond on morfosüntaks, uurimisobjekt keelekasutusmustrid. Aluseks on konstruktsioonigrammatika (Croft 2003; Goldberg 2006) ja funktsionaalse grammatika (Praha lingvistiline koolkond, vene keele grammatikateooria) sidusus kasutuspõhises keelekäsitluses (nt Barlow, Kemmer 2000: 7–28; Bybee 2007; 2010; Bybee, Hopper 2001; Verhagen 2009; Ellis jt 2013). Võrreldud on õppija ja emakeelekõneleja keelekasutusmustreid, nende süntaktilise ja leksikaalse varieerumise piire verbist vasakule jäävas kontekstis (Eslon 2013; 2014); kirjeldatud verbist paremale jäävat konteksti ja objektifraasi grammatikat ning semantikat (Eslon 2012; Eslon, Õim 2010); käänete tekstiparadigmat ja käändeasendusi (Eslon 2008; 2009; 2010a; 2010b; 2011; Eslon, Matsak 2009); eitust (Kitsnik 2007); sõnajärge (Metslang, Matsak 2010; Kaivapalu 2010; Matsak jt 2010a; Matsak jt 2010b). EVKK tekste on kasutatud teise keele omandamisprotsessi uurimiseks ja kirjeldamiseks lähisugulaskeeltes (Kaivapalu, Eslon 2011; Eslon jt 2010; Kaivapalu 2009; Eslon, Kesksaar 2009; Kaivapalu 2008; Eslon 2006; 2007; Kitsnik 2006).

Keelekasutusmustrid (mitte sõnad, vormid või vealiigid) on teadlik valik, sest mitmest komponendist koosnevates ning regulaarselt kasutatud keelestruktuurides tulevad esile sõna- ja vormivaliku piirangud ning nende leksikaalse ja morfosüntaktilise varieerumise piirid (vt Eslon 2013; 2014). Mitmesõnalistel kooslustel on kindel koht nii õppija kui ka emakeelekõneleja tekstiloomes. Kuna vormide ja sõnade kooskasutuse printsiibid on inkorporeeritud aktuaalsesse grammatikasse ja seetõttu universaalne nähtus (Prince, Smolensky 2002), siis on kasutusmustrid sobiv alus nii emakeelekõneleja ja õppija kirjaliku tekstiloome võrdlemiseks kui ka lugemis-, rääkimis- ja kirjutamisoskuse kujundamiseks. Ühisjoonte ja sarnasuse kõrval tuleb esile ka suuri erinevusi: kasutatakse sama keelt, samu sõnu ja vorme, kuid neid kombineeritakse erinevalt; ka nende sagedus ja osakaal on keelekasutuses erinev (vt Eslon 2013; 2014).

Teiseks annavad keelekasutusmustrid olulist teavet keeleõppe materjalide koostajatele, õpetajatele ja (ise)õppijatele. Korpusanalüüsi tulemusel leitud keelekasutusmustrite võrdlus üldkasutuses olevate õpikute sõnavara ja grammatikateemadega näitab lõhet tegeliku ja õpetatava keelekasutuse vahel, reaalselt toimivate kasutusreeglite ja grammatikakirjeldustes sisalduvate reeglite vahel. Kasutuses olulised reeglid ja õpikute grammatikareeglid on tihtilugu ka teistes keeltes vastukäivad, mida on osutanud nt Ute Römer (2007) seoses inglise keele *if*-lausete õpetamisega. Amy B. M. Tsui küsitles õpetajaid kuueteistkümne leksikaalgrammatilise kategooria õpetamiseks kasutatud õppematerjalide relevantsuse kohta ning sai pigem negatiivse vastuse. Järelikult tuleks õpetajatel ja õppematerjalide koostajatel grammatikate asemel tugineda korpusandmetele (Tsui 2005: 336, 338) ja kasutusgrammatikale.

Kolmandaks on eesti kirjakeele ja õppijakeele kasutuspõhise võrdleva analüüsi tulemustest kasu õppijakeele automaatse analüüsi rakenduste loomisel. Olenemata süsteemiarenduse lähtekohtadest, on keskne küsimus ikkagi lingvistiline: kuidas sõnu, vorme ja keelestruktuure tüüpiliselt kombineeritakse ehk mis ühendab sõna semantikat, vormi (grammatika) ja vormi tekstilisi funktsioone. Klastrileidja otsib tekstidest üles lingvistilised tervikobjektid,

mis kasutuses sagedad ja seetõttu teksti mõistmiseks ning produtseerimiseks olulised (vt ka Conklin, Schmitt 2008). Nende tervikobjektide morfosüntaktilise ja leksikaalse varieerumise põhjal saab leida keeleüksuste lineaarse kooskasutuse piirangud ehk reeglid, mida grammatika ei kajasta. Seetõttu on nendel piiratud sõna- ja vormivalikuga tekstilistel struktuuridel ühelt poolt kindel koht lugemisoskuse arendamisel, rääkimis- ja kirjutamisoskuse kujundamisel ning teisalt teksti automaatseks analüüsiks ja sünteesiks vajalike vahendite loomisel.

4. Kokkuvõte

Ülevaade Tallinna Ülikooli eesti vahekeele korpusest annab ettekujutuse korpuse esimese versiooni kasutusvõimalustest empiirilises ja rakenduslikus uurimistöös, tulevaste õpetajate ja lingvistide koolitamises ning täiendõppes. Korpuskeskkonnas saab teostada eesti-, vene- ja soomekeelsete tekstide automaatset morfoanalüüsi, leida tekstikasutuses kaks või enam korda esinenud morfoloogilisi, süntaktilisi ja morfosüntaktilisi terviküksusi, kasutada sagedus- ning pöördsõnastikku, silbitajat, leida lemmasid, saada ülevaate liidetest ja liitelisest tuletusest, leida tekstides sisalduvaid illustreerivaid (vea)näiteid jm. Kombineerides erinevaid tekstilisi tunnuseid (nt tekstiliik, sõnade arv, lausete hulk), vealiike (hierarhiline taksonoomia) ja metateavet õppija kohta (nt emakeel, päritolumaa, sugu, haridus), võimaldab EVKK kasutajaliides mitmetasandilist otsingut. Korpusse saab importida mis tahes tekste ja neid seal analüüsida; paralleelne võimalus on eksportida korpuse tekste ja rakendada õppija keelekasutuse uurimiseks meetodeid, mida korpuskeskkond ei paku.

EVKK esimene versioon on aluseks loodavale eesti õppijakorpusele – eesti keele kui riigikeele tasemeoskuse testimise keskkonnale, kuhu igaüks võib soovi korral sisestada teksti ning saada tagasisidet. See on mugav kõigile, kes soovivad keelt õppida omas tempos ja endale sobival moel. Õppijakorpuse juurde kuuluksid veebiteenused, mis võimaldavad integreerida vealeidja ja keeleoskustaseme määramise erinevatesse veebipõhistesse õpikeskkondadesse (Moodle, Dippler). Sel viisil lisanduksid e-õppeplatvormile nii keeletaseme omandamiseks vajalikud kontrollharjutused kui ka enesekontrollitestid. Niisuguse keskkonna loomisele eelneb iga keeleoskustaseme lingvistiline ja statistiline modelleerimine, mudelite evalveerimine ning rakenduste loomine, õppijakorpuse kontseptuaalse disaini väljatöötamine, automaatne veaotsing ja vealiigi identifitseerimine, reeglite väljastamine koos viidetega sõnastikele ja kasutusgrammatikale. Niisiis seisab ees hulgaliselt lingvistilist ja tehnoloogilist tööd, mis nõuab aega ja vahendeid.

EVKK loomist ja arendamist on toetanud riikliku programmi „Eesti keele keeletehnoloogiline tugi (2006–2010)" projekt „VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine (2008–2010)" ja „Eesti keel ja kultuurimälu (2009–2013)" projekt „REKKi käsikirjaliste materjalide digiteerimine, Eesti vahekeele korpuse alamkorpuste loomine ja korpuse kasutusvõimaluste populariseerimine (2009–2013)".

Kirjandus

A n t h o n y, Laurence 2013. A critical look at software tools in corpus linguistics. – Linguistic Research, kd 30, nr 2, lk 141–161.

B a r l o w, Michael, K e m m e r, Suzanne (toim) 2000. Usage-Based Models of Language. Stanford: CSLI.

B y b e e, Joan 2007. Frequency of Use and the Organization of Language. Oxford: Oxford University Press.

B y b e e, Joan 2010. Language, Usage and Cognition. Cambridge: Cambridge University Press.

B y b e e, Joan, H o p p e r, Paul (toim) 2001. Frequency and the Emergence of Linguistic Structure. Amsterdam: John Benjamins.

C o n k l i n, Kathy, S c h m i t t, Norbert 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? – Applied Linguistics, kd 29, nr 1, lk 72–89.

C r o f t, William 2003. Typology and Universals. Second edition. Cambridge: Cambridge University Press.

E l l i s, Nick C., O ' D o n n e l l, Matthew B., R ö m e r, Ute 2013. Usage-based language: Investigating the latent structures that underpin acquisition. – Language Learning, kd 63, nr s1, lk 25–51. http://www.uteroemer.com/Language%20Learning%20VACs%20Ellis%20et%20al%202013.pdf (20. V 2014).

E s l o n, Pille 2006. Eesti vahekeele korpusest korrelatsioonigrammatikani. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 2, lk 11–24.

E s l o n, Pille 2007. Õppijakeelekorpused ja keeleõpe. – Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine. (Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 9.) Tallinn: Tallinna Ülikooli Kirjastus, lk 87–120.

E s l o n, Pille 2008. Käändevormide kasutussageduse võrdlus eesti õppijakeeles ja kirjakeeles. – Õppijakeele analüüs: võimalused, probleemid, vajadused. (Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 10.) Tallinn: Tallinna Ülikooli Kirjastus, lk 31–66.

E s l o n, Pille 2009. Eestikeelses tekstiloomes eelistatud konstruktsioonid ja käändevormid. – Korpusuuringute metodoloogia ja märgendamise probleemid. (Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 11.) Tallinn: Tallinna Ülikooli Kirjastus, lk 30–53.

E s l o n, Pille 2010a. Suundumustest eesti keele grammatiliste käänete kasutamisel. – Korpusuuring ja meetodid. (Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 12.) Tallinn: TLÜ EKKI, lk 7–36.

E s l o n, Pille 2010b. Muutustest eesti keele grammatiliste käänete kasutamisel. – Lähivõrdlusi. Lähivertailuja, nr 19, lk 38–60.

E s l o n, Pille 2011. Millest räägivad eesti keele käändeasendused. – Lähivõrdlusi. Lähivertailuja, nr 21, lk 45–64.

E s l o n, Pille 2012. Objekti ja tegevuse markeeritus eesti õppijakeeles. – Lähivõrdlusi. Lähivertailuja, nr 22, lk 15–42.

E s l o n, Pille 2013. Kahe keelekasutusvariandi võrdlus: morfoloogilised klassid ja klastrid. – Lähivõrdlusi. Lähivertailuja, nr 23, lk 13–38.

E s l o n, Pille 2014. Morfosüntaktilise ja leksikaalse varieerumise piiridest: ilukirjandus- ja õppijakeele kasutusmustrite võrdlus. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 10, lk 55–71.

E s l o n, Pille, K e s k s a a r, Anneli 2009. Keeleõpe arvuti abil. – Tiigriõpe: Haridustehnoloogia käsiraamat. Tallinn: Tiigrihüppe sihtasutus, lk 117–132.
E s l o n, Pille, M a t s a k, Erika 2009. Eesti keele kasutusvariandid: korpusest tulenev käändevormide võrdlev analüüs. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 5, lk 79–110.
E s l o n, Pille, M e t s l a n g, Helena 2007. Õppijakeel ja eesti vahekeele korpus. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 3, lk 99–116.
E s l o n, Pille, Õ i m, Katre 2010. Objektikäänete kasutamisest sageduse ja markeerituse seisukohalt. – ESUKA, nr 1-2, lk 69–89.
E s l o n, Pille, Õ i m, Katre, K a i v a p a l u, Annekatrin, A r g u s, Reili, M a t s a k, Erika 2010. Kuidas uurida esimese ja teise keele omandamist? – Lähivõrdlusi. Lähivertailuja, nr 20, lk 11–48.
G o l d b e r g, Adele E. 2006. Constructions at Work. The Nature of Generalization in Language. Oxford: Oxford University Press.
K a i v a p a l u, Annekatrin 2008. Lähtekeele mõju korpuspõhine uurimine. – Õppijakeele analüüs: võimalused, probleemid, vajadused. (Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 10.) Tallinn: Tallinna Ülikooli kirjastus, lk 93–119.
K a i v a p a l u, Annekatrin 2009. Õppijakeele korpusanalüüsi täiendavatest meetoditest. – Korpusuuringute metodoloogia ja märgendamise probleemid. (Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 11.) Tallinn: Tallinna Ülikooli Kirjastus, lk 72–98.
K a i v a p a l u, Annekatrin 2010. Mõnede eesti sõnajärjemallide psühholingvistilisest reaalsusest. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 6, lk 103–120.
K a i v a p a l u, Annekatrin, E s l o n, Pille 2011. Onko lähisukukielen vaikutus suomen ja viron omaksumiseen symmetristä? Korpuspohjaisen tutkimuksen tuloksia ja haasteita. – Lähivõrdlusi. Lähivertailuja, nr 21, lk 132–153.
K a l l a s, Jelena, K o p p e l, Kristina, T i i t s, Mai, T u u l i k, Maria (toim) 2014. Eesti keele põhisõnavara sõnastik. Tallinn: Eesti Keele Sihtasutus.
K i t s n i k, Mare 2006. Keelekorpused ja võõrkeeleõpe. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 2, lk 93–107.
K i t s n i k, Mare 2007. Õppijakeele uurimise ja arendamise võimalusi eesti vahekeele korpuse põhjal (eituse väljendamise näitel). Magistritöö. Käsikiri Tallinna Ülikooli eesti keele ja kultuuri instituudis.
К о п о т е в, Михаил, М у с т а й о к и, Арто 2003. Принципы создания Хельсинкского аннотированного корпуса русских текстов (ХАНКО) в сети Интернет. – Научно-техническая информация, серия 2, № 6, lk 33–37.
L õ o, Kaidi 2012. The relationship between lexical richness and Estonian learners language proficiency levels. Seminaritöö käsikiri. Seminar für Sprachwissenschaft. Eberhard Karls Universität Tübingen.
M a t s a k, Erika, E s l o n, Pille, K i p p a r, Jaagup 2010a. Eesti keele sõnajärje vealeidja prototüübi arendamine. – Korpusuuring ja meetodid. (Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 12.) Tallinn: TLÜ EKKI, lk 59–100.
M a t s a k, Erika, M e t s l a n g, Helena, K i p p a r, Jaagup 2010b. The prototype of word order assessment at the Estonian Interlanguage Corpus. – The 2010 International Conference on Artificial Intelligence. Las Vegas: CSREA Press, lk 870–875.

Metslang, Helena, Matsak, Erika 2010. Kesksete lausekomponentide järjestus õppijakeeles: arvutianalüüsi katse. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 6, lk 175–193.

Meurer, Paul 2012. Corpuscle – a new corpus management platform for annotated corpora. – Exploring Newspaper Language: Using the web to create and investigate a large corpus of modern Norwegian. (Studies in Corpus Linguistics 49.) Toim G. Andersen. Amsterdam: John Benjamins, lk 29–50.

Ots, Sander 2012. Statistikapõhise tarkvara loomine morfoloogiliste kollokatsioonide eraldamiseks eesti keele tekstidest. Bakalaureusetöö. Käsikiri TLÜ informaatikainstituudis.

Pajupuu, Hille 2007. Kuidas hinnata suure panusega testide hindajaid. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 3, lk 221–233.

Pepper, Steve 2012. Lexical transfer in Norwegian interlanguage. A detection-based approach. Master Thesis in Linguistics. University of Oslo. https://www.duo.uio.no/bitstream/handle/10852/34792/MasteroppgavexxPepperx.pdf?sequence=1 (20. V 2014).

Prince, Alan, Smolensky, Paul 2002. Optimality Theory. Constraint Interaction in Generative Grammar. http://roa.rutgers.edu/files/537-0802/537-0802-PRINCE-0-0.PDF (20. V 2014).

Rannut, Mart 2008. Estonianization efforts post-independence. – International Journal of Bilingual Education and Bilingualism, kd 11, nr 3/4, lk 423–439.

Römer, Ute 2007. Learner language and the norms in native corpora and EFL teaching materials: A case study of English conditionals. – Anglistentag 2006 Halle. Proceedings. Trier: Wissenschaftlicher Verlag Trier, lk 355–363.

Sirts, Kairit 2012. Noisy-channel spelling correction models for Estonian learner language corpus lemmatisation. – Human Language Technologies. The Baltic Perspective: Proceedings of the Fifth International Conference Baltic HLT 2012. (Frontiers in Artificial Intelligence and Applications 247.) Toim A. Tavast, K. Muischnek, M. Koit. Amsterdam: IOS Press, lk 213–220.

Trainis, Jekaterina 2013. New view to Estonian literary language: Cluster analysis and its application. – Linguistics Beyond and Within. International Linguistics Conference in Lublin 14–16 November 2013. Book of Abstracts. Lublin, lk 107–108. http://lingbaw2013.webclass.co/schedule/Book%20of%20abstracts.pdf (20. V 2014).

Trainis, Jekaterina, Allkivi, Kais 2014. Ilukirjanduskeelest uue pilguga. – Eesti Rakenduslingvistika Ühingu aastaraamat, nr 10, lk 283–306.

Tsui, Amy B. M. 2005. ESL teachers' questions and corpus evidence. – International Journal of Corpus Linguistics, kd 10, nr 3, lk 335–356.

Vajjala, Sowmya, Lõo, Kaidi 2013. Role of Morpho-Syntactic Features in Estonian Proficiency Classification. – Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications (BEA8). Atlanta: Association for Computational Linguistics, lk 63–72.

Verhagen, Arie 2009. The conception of constructions as complex signs. Emergence of structure and reduction to usage. – Constructions and Frames, nr 1, lk 119–152.

Estonian Interlanguage Corpus

Keywords: multilevel search, statistics, annotation, linguistic error taxonomy, Estonian learner language, morphosyntactic patterns of language usage

The article introduces the first version of the Estonian Interlanguage Corpus (EIC) of Tallinn University, surveying the corpus structure, multilevel statistics, corpus annotation, linguistic error taxonomy, system of requesting, options of automatic analysis (morphological and syntactic analysis, n-grams) of Estonian learner language, and current EIC-based research.

EIC is a resource consisting of Estonian texts written by learners of Estonian as an official and foreign language. The corpus has hitherto provided material for empirical and applied research on morphosyntactic usage patterns and lexical variation of Estonian, the morphosyntactic complexity and lexical richness of learner language, developments in the Estonian language system, gradual development of language skills and CEFR proficiency levels, error and contrastive analysis (Estonian, Russian, Finnish morphology), and cluster analysis. EIC is a monitor corpus containing over three million word forms.

The major direction of research is comparative corpus analysis of standard Estonian and learner Estonian in the domain of morphosyntax, the focus lying on patterns of language usage. This has been a conscious choice as the multi-component language structures which are regularly used have a definite place in the text creation process of a native speaker as well as a learner, while the comparative analysis of the resulting texts has a heuristic meaning for understanding Estonian grammar. The results can also benefit applications for automatic analysis of learner language. Regardless of the starting points of system development, the central issue will still be linguistic, namely, how do people typically combine words, morphology and linguistic structures or, in other words, what is the connection between the semantics and morphology of a word and the textual functions of its morphology.

Pille Eslon (b. 1950), PhD, Tallinn University, Institute of Estonian Language and Culture, associate professor, pille.eslon@tlu.ee