ÕPPIJAKEELE-KORPUSED JA KEELEÕPE

Pille Eslon

Ülevaade

Artiklis käsitletakse praegusaja keeleõppe võimalusi ja vajadusi. Tähelepanu keskmes on õppijakeelekorpused, nende arendamine teaduslikel ja kommertseesmärkidel, kasutamine teise keele / võõrkeeleõppe õppimise ja õpetamise edendamiseks. Loodud on keeleõppesüsteeme, mis lubavad rakendada arvutipõhist keeleõpet, tööle panna elektroonilised grammatikad ja sõnastikud. Artiklist leiab lugeja ülevaate selle kohta, milliseid võimalusi pakub *Eesti vahekeele korpus.* Konkreetsemalt peatutakse veamärgendusel ja sõnasagedusel[1].

Võtmesõnad: korpused ja keeleõpe, Eesti vahekeele korpus, vigade märgendamine, taksonoomia ja statistika

Mis on õppijakeel ja õppijakeelekorpus?

Mõisted *õppijakeel* ja *vahekeel* on kujunenud erinevates paradigmades, kuid üldjoontes tähistatakse mõlemaga keelevariante,

[1] Eesti vahekeele korpuse loomine ja arendamine on toimunud järgmiste projektide toel: sihtfinantseeritav teema nr 0132493s03 "Eesti keelekeskkonna arengu analüüs, modelleerimine ja juhtimine" (2003–2007); riiklik programm "Eesti keel ja rahvuslik mälu" (2004–2008), grant R 05/01 "Koodivahetuse, vahe- ja lastekeele korpuste töötlemine ja haldamine"; ETFi grant nr 6151 "Koodivahetuse, eesti vahekeele ning lastekeele andmekorpuste koostamine ja üldkirjeldus" (2005–2008).

mida õppijad sihtkeeles loovad. Termini *vahekeel* (*interlanguage*) võttis 1972. aastal kasutusele L. Selinker (1992: 31–54). Selleks lõid soodsa pinnase biheivioristlik keelekäsitlus, keelte kontrastiivanalüüs (lähte- ja sihtkeele võrdlus) ning arenev interferentsiteooria (lähtekeele negatiivne mõju sihtkeelele). Terminit *õppijakeel* (*learner language*) on kasutatud seoses teise keele / võõrkeele omandamisega (*second / foreign language aquisition*). Keskse kohal on keelevea (*error*) mõiste, veaanalüüs (*error analysis*) ja veataksonoomia (vt Corder 1981). Vea mõiste täpsustamine toimub võrdluses autentse sihtkeelega, mille tulemusel hakatakse eristama keelelise kreatiivsuse ilminguid ja väsimusest jne tulenevaid eksimusi. Lähtekeele ja õppijakeele võrdluse tulemused ei piirdu negatiivse ülekande konstateerimisega – kõrvuti interferentsiga märgatakse ka võimalikku positiivset mõju sihtkeele omandamisele. Suhtlemiseks piisava keelekompetentsi kujunemine toimub astmeliselt, õpitakse enese ja teiste vigadest, protsessi suunavad erinevad vajadused, suhtlus-, õpi- ja õpetamisstrateegiad (Kaivapalu 2006: 72–74, 83–85).

Õppijakeelekorpus (*learner corpus, learner corpora*) kuulub kaasaegsete elektrooniliste keeleressursside alla. 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, korpusuuringutele tuginevad oma töös teise keele / võõrkeele õpetamise spetsialistid (Granger 2003: 465). Õppijakeele võrdlemine autentse kirjakeelega tõlke- ja paralleelkorpuste andmeid kaasates võib anda üllatavaid tulemusi, mis osutavad alles kujunema hakkavatele nähtustele ja uutele arengutele sihtkeele grammatikas või leksikaalgrammatilises perifeerias.

Õppijakeelekorpustele tuginedes on tekkinud *arvutipõhine keeleõpe* (*computer aided language learning*), nt prantsuse sihtkeelega õppijakeelekorpuse FRIDA (*French Interlanguage Database*) alusel on loodud *FreeText*, tänu millele on tekkinud reaalne võimalus siduda keeleõpe keele omandamisega.

Kuidas korpusi keeleõppes kasutatakse?

Korpuste kasutamisel keeleõppes võib olla erinevaid võimalusi: autentse keeleainese allikast automatiseeritud interaktiivse õpikeskkonnani. Selleks otstarbeks on sobivad nii veebikeskkond enam kui biljonilise dokumendihulgaga (Volk 2002: 355) kui ka kirjakeele, paralleel-, tõlke-, keeleõppe tekstide, õppijakeele- ja õppijakorpused. Vajalikuks on peetud mitte lahutada kirjalikku ja suulist kõnet, kuna loomulikus keelekasutuses on mõlemad omavahel orgaaniliselt seotud (vt Myles 2005: 375). BNC (*British National Corpus*) tegijad on pidanud ideaalseks, kui korpuses on võrdselt nii kirjalikke tekste kui suulist kõnet, olgugi et tegelikult pole see neilgi reaalseks osutunud (BNC sisaldab 10% suulise keelekasutuse materjale ja 90% kirjalikke tekste). Nagu lapsekeele andmestikku, nii on ka suulist ja kirjalikku (õppija)keelt raske koguda, ühtse standardi alla viia ja omavahel siduda. Pealegi on transkribeerimine mahukas ning aeganõudev töö, mida ei lähe vaja sugugi igas teise keele / võõrkeelega seotud uurimuses (Granger 2004 : 124–125, 130).

1. Kirjakeele, tõlke- ja paralleelkorpused, veebikeskkond

Keeleõppe arendamise eesmärgil on **kirjakeele korpus(t)est saadavate andmete lingvistiline analüüs** oluline uut laadi sõnastike ja õppematerjalide koostamiseks, samuti ainekavade sisu-

liseks täiendamiseks materjalidega, mida emakeelekõneleja igapäevasuhtluses aktiivselt kasutab. Eesti keelega seoses võib näiteks tuua A. Kilgi (Kilgi 2005) lühiuurimuse tulemused verbisageduse muutumisest ainsuse kolmanda pöörde vormis. Aluseks on võetud Tartu Ülikooli kirjakeele korpuse alamkorpused vahemikus 1890 kuni 1970 ja vaadeldud, milliste verbidega saab seostada muutusi ainsuse kolmanda pöörde vormi sageduses. Kui 1930. aastate korpuses kasutatakse aktiivselt vorme *toimub, esineb, kasutab, areneb, märgib, ületab, võimaldab, omab*, siis 1950. aastate korpuses verbe *töötab, areneb, märgib, täidab, võimaldab, ületab, sammub, omab* ning harvaks on jäänud või üldse ei esine sõned *näib, arvab, juhtub, seletab, sureb, kõlbab, julgeb*. Nimetatud perioodi vältel esines kõige sagedamini oleviku vorm *on*, millele järgnesid modaalid *tuleb, võib, peab, saab* (Kilgi 2005: 4). Kirjeldatud faktid on olulised keeleõppe ainese valikul. Asi pole pelgalt selles, et keelt õppides tuleb keskenduda vormimoodustusele või verbide *olema, tulema, võima, pidama, saama* omandamisele, vaid selles, et nende verbide vorme kasutatakse erineva sagedusega ning kindlates verbitarindites ja verbikesksetes lausestruktuurides. See omakorda sõltub teatud mõistete ja tähenduste edastamisvajadusest ja kajastub eelkõige verbide rektsioonistruktuuride erinevustes. Nii on modaalide *tuleb, võib, peab, saab* kasutusvaldkond eelkõige infinitiivtarindid (*tuleb teha, võib teha, peab tegema, saab teha*), kus finiitsele modaalile järgneva infinitiivi vormi valik oleneb modaali semantikast, tarindi kasutus aga kõneakti pragmaatikast. *Olema*-verbi ainsuse kolmandat pööret kasutatakse enamasti koopula funktsioonis, järgnev öeldistäide on piiritletud kindlate sõnaliikide ning grammatiliste vormidega, kindlate süntaktiliste struktuuridega jne. Sedalaadi seoste nägemine on oluline eesti keele kui teise keele / võõrkeele ainekava sisu lahtimõtestamisel, kõnearendusteemade ja õppetekstide valikul, järjestamisel ning adapteerimisel. Tavaliselt tuginetakse siin vea- ja kontrastivanalüüsi tulemustele, kuid enamasti

siiski intuitsioonile, praktilisele õpetamiskogemusele ning aja jooksul kujunenud traditsioonidele. Mõttekas oleks aga teada, mis reaalses keelekasutuses tegelikult toimub. Seetõttu ongi keeleõppe tarvis väärtuslikud need uuringud, mis tuginevad suurtele lingvistilistele andmekogudele ning täheldavad muutusi keelekasutuses (sõna- ja vormivariatiivsus, sõnade ja vormide kasutussagedus, kollokatsiooniliste struktuuride esinemine jne). Näiteks 1990. aastate ajakirjanduse ja televisiooniesinemiste keelepruugi analüüsi põhjal on H. Metslang järeldanud, et verbitarindi *hakkan laulma* kasutamine tuleviku tähenduses on aina laialdasemalt leviv nähtus, mis on leidnud analoogiat ka karjala, ungari ja vepsa keeles (Metslang 1994: 216–218). Põhjus on eesti keele sisene: "eesti oleviku vorm on markeerimata ja sobib viitamiseks nii olevikulisele, tulevikulisele, üldkehtivale ajavälisele kui ka minevikulisele, nii ühekordsele kui ka korduvale, kestvale kui ka hetkelisele sündmusele" (Metslang 1994: 534). Kordusuuringus 2004–2005 tugines H. Metslang juba eesti veebikeele analüüsile. Selgus, et *hakkama*-tarindiga viidatakse tulevikule üsnagi laialt ja kuigi kasutusalas muutusi pole märgata, on siiski eri eluvaldkondade temaatika osas välja kujunenud mõned tüüpkontekstid: *Senise energeetikaosakonna laenudega hakkab tegelema linnavalitsuse rahandusosakond*; *Neid autosid hakkab müüma selleks otstarbeks loodud omaette firma* (Metslang 2005: 7–8). Järelikult on eesti keele õppe seisukohalt oluline mõista, et *hakkama*-tarindi kasutus on seotud kaugema tulevikulisuse väljendamisega ja kindlate lausestruktuuridega, mis omakorda piiravad suhtlussfääri ja kommunikatiivse registri valikut.

Analoogseid eesti keele uurimusi on tänaseks rohkesti – tõsiselt pole võetav töö, mis ei tugine suuremahulisele andmebaasile ega ole töödeldud erinevate arvutiprogrammide ja kvantitatiivseid / kvalitatiivseid analüüsimeetodeid kasutades. Aktiivselt on uuritud ajakirjanduskeele spetsiifikat ning arenguten-

dentse. Näiteks R. Kasik peab ajakirjanduskeele üheks oluliseks funktsiooniks uue sõnavara sissetoomist ja sellega kaasnevaid protsesse (Kasik 2004: 5–14), K.Kerge on jälginud lause süntaktilise keerukuse ajalisi muutusi (Kerge 2003), eestlase viisakusruumi (Kerge 2006: 67–89) jne.

Mujal maailmas on korpuspõhiseid lingvistilisi uuringuid peetud keeleõppe eesmärgil väga oluliseks (vt Kitsnik 2006: 94–99). Eelduse selleks on loonud keeletehnoloogilised rakendused, tänu millele on korpuste tekstid morfoloogiliselt, süntaktiliselt ja semantiliselt ühestatud, maha on võetud homonüümia. Korpustes saab teha erinevaid päringuid, töödelda andmeid statistilistelt jms. Korpuspõhisest keeleõppest on välja kujunenud omaette suund: korpuste alusel luuakse uusi üheja mitmekeelseid ning sagedussõnastikke, leitakse tihti kasutatavaid kollokatsioone, koostatakse autentseid õppematerjale, modelleeritakse tasemeõpet ja koostatakse teste. Samas on kirjakeele korpuste kasutamisega mindud ka äärmusse, kuna korpustes nähakse keeleõppe ainukest allikat ja imerohtu, mis peaks aitama lahendada kõik probleemid. Nagu äärmustega ikka, pole ka see seisukoht leidnud piisavat toetust – näiteks segaduste tõttu, mille võib esile kutsuda mõistepaar *kasutussagedus* ja *produktiivsus*, sest sõnade kasutussageduse, vormimoodusparadigmade ning nende produktiivsuse vahel ei ole üksüheseid ja samasuunalisi seoseid. A. Nikolajevi uurimused soome keele muutkondade produktiivsuse indeksi määramisel on viinud autori tõdemuseni, et sõnatuletusmalli aktiivsuse ja muutkondade produktiivsuse vahel on hajusad seosed. Sellele järeldusele oli võimalik jõuda tänu lingvistiliselt töödeldud suurte andmebaaside kvantitatiivsele analüüsile – A. Nikolajev tugines 130 miljonist sõnest koosnevale soome keelepangale[2] ja tegi selle alamkorpustes üle 75000 päringu, mille tulemusena

[2] http://www.csc.fi, 6.07.07

leidis 49 nimisõna muuteparadigmat ning võrdles neid omavahel (Nikolajev 2007: 74).

Mis puudutab tänapäeva eesti keele korpuspõhiseid uurimusi, siis tuleb märkida, et nende tulemusi pole seni keeleõppe eesmärgil üldistatud ega keeleõppe jaoks olulist lingvistilist teavet välja toodud. Samas on loomulik, et ainekava koostajad tugineksid oma töös näiteks eesti keele sagedussõnastiku andmetele (Kaalep & Muischnek 2002), õpiku autorid aga leiaksid vajalikku keeleainest kollokatsioonide ja püsiühendite andmebaasist (vt Muischnek 2006) jne.

Teine suund korpuspõhises keeleõppes on seotud **paralleel- ja tõlkekorpuste kontrastiivse ning tõlkeanalüüsiga**. Eesmärk on leida tõlkimise ja õppijakeele universaale (nt Mauranen & Kujamäki 2004), tüüpilisi leksikaalseid ja süntaktilisi vastavusi / mittevastavusi erinevate keelte vahel, et välja töötada optimaalne tõlkijakoolituse ja keeleõppe mudel, veenduda sõnastike ja keeleõppematerjalide sobivuses tõlkija ning keeleõppija vajadustega, uurida strateegiaid, mida inimene tekstide loomisel ja tõlkimisel kasutab. Seejuures annab korpuslingvistiline tõlkeuurimus meetodi, mille abil pääseda tõlke olemuseni; süstemaatilise tõlkeanalüüsi rakendamine annab aga keelte kõrvutamiseks sobiva meetodi, mis aitab viia kontrastiivse keeleuurimise kvalitatiivselt uuele tasandile (vt Jantunen & Eskola 2002 : 202; Eslon 2006a: 17, 19–20).

Paralleel- ja tõlkekorpuste alusel saab jälgida ka kirjakeele arengusuundi. Selleks on vaja kõigepealt lahendada normatiivsuse küsimus ja võrrelda tõlget nii kirjakeele korpuste samaväärsete andmetega kui tõlkekorpustes sisalduvate lähtekeele tekstide tõlgetega erinevatesse sihtkeeltesse. Nii saab vähendada või vältida võimalikku keele- ja kultuuriinterferentsi, valesid ja ebatäpseid tõlkevasteid, kalkeerimist jms (vt Михайлов 2003: 31; McEnery & Wilson 2001: 72), suunata tõlkestrateegia

valikut, et vältida sihtkeele vahendite üle- ja alakasutust, ebatüüpilisi leksikaalseid ja süntaktilisi kooslusi (vt Jantunen 2007b). Näiteks võiks tuua Joensuu ülikooli 10 miljonist sõnest koosneva tõlkekorpuse, mis sisaldab seitset liiki tekste akadeemilistest meelelahutuslikeni ja mille lähtekeeled on inglise, vene, prantsuse, rootsi, saksa, ungari ja eesti (A. Mauraneni projekt "Käännössuomi ja kääntämisen universaalit. Tutkimus korpusainestolla"), samuti Jyväskylä paralleelkorpuse (K. Sajavaara), mida tehti koostöös Oslo inglise-norra korpusega (S. Johansson). Paralleelkorpusesse oli seega peale sugulaskeelte haaratud ka soome kui mittesugulaskeel.

Sisuliselt on tõlge samaväärne mistahes muu keelevariandiga ning selles mõttes võib näha analoogiat õppijakeele tekstide võrdluses kirjakeele korpuse tekstide või veebikeelega – kõik nad on ühe keele paralleelvariandid. Seda protsessi võib pidada ka keelesisese tõlke ilminguks. Küsitavusi tekitab aga nii õppijakeele kui tõlketekstide analüüsimine eraldi lausete kaupa, kui neid hinnatakse õige / vale skaalal, võttes võrdlusaluseks uurija koostatud "õige" lause. Niisugune lähenemine on ühekülgne, kuna pole arvestatud tervikteksti ja selle autoriga. Samuti on määramata, milles seisneb erinevus mõistete *normatiivne*, *õige* ja *vale* vahel.

Siinkohal lisagem, et eesti keele paralleel- ja tõlkekorpuste loomisel tehakse tõsist tööd (nt Uibo & Bick 2005; Uibo 2006).

Kolmas suund korpuspõhises keeleõppes on seotud **veebikeele uurimistulemustega**. Google´i materjalidest võib koostada väiksemaid, spetsiifilisi vajadusi rahuldavaid teksti- või näidetekogusid, mida saab vastavalt eesmärgile adapteerida, mugandada, rakendada ülesannetes, võtta aluseks õigekirja küsimustes jne. Samas võime kasutada ka kogu Google´is kättesaadavat üle biljonilist tekstipanka, et uurida, millised protsessid elavas keelekasutuses reaalselt toimivad (nt aktiivsed liitsõnade

moodustusmallid, uue sõnavara tekkimine või sõnalaenude vohamine ja mugandumine, tavapärased või ebaharilikud kollokatsioonid, vormi- ja struktuurieelistused, autorikeele omapära jne), millest inimesed huvituvad või mis on ühiskonnas aktuaalne. Niisugust materjali ei leia klassikalistest kirjakeele standardkorpustest nagu *British National Corpus* (BNC) ja vene kirjakeele korpus *Ruscorpora* või *Tartu Ülikooli eesti kirjakeele korpus* (TÜKK). Erandiks on ajakirjanduskeele *online*-korpused (sh ka *Eesti Keele Instituudi eesti kirjakeele korpus*), mis on nii info kui keelekasutuse poolest standardiseerimata ja heterogeensed. Näiteks G. Bergh on võrrelnud sõnaotsingu "Taliban" tulemusi Coubildi, BNC ja Google´i alusel. Tulemus hämmastas uurijat ennastki: Coubildi 56 miljoni sõne seast leidis ta 40 konteksti, BNC 100 miljonist – mitte ühtegi ja Google´i ingliskeelsest variandist – 1 890 000 näidet (Bergh 2005: 26–27). Analoogselt võrdles ta variantide *colour* ja *color* kasutussagedust standardkorpustes (Browni korpus, LOB, BNC) ja Google´is, eristades ameerika ja briti veebi inglise keelt (Bergh 2005: 39–42). Kahtlemata on veebikeele uurimise eelis representatiivsema tulemuse saamine: selguvad antud ajahetkel keelele omased, tüüpilised, kesksed nähtused ja protsessid. Sellest ei saa oma töös mööda minna ka keelekorraldajad. Veebikeele uuringute läbiviimise probleem seisneb aga piisavalt kasutajasõbraliku vabavara leidumises.

Neljas korpuspõhise keeleõppe arendamise suund ongi olnud seotud **erineva korpustarkvara, standardiseeritud arvutiprogrammide** (nt Oxford Concordance Program, Word Cruncher, WordSmith Tools, MonoConc, Text Encoding Initiative – TEI jne) **ning statistiliste meetodite rakendamisega keelevara, keelevariatiivsuse, keelemuutuste ja keelekogukondade uurimisel** (nt Johnson 2006, Oakes 1998). Mõõtes grammatilisi vorme algebraliste suurustega, saame võimaluse minna süvitsi keele funktsioneerimise seaduspärasuste tõlgendamisse, võrrelda

uue nurga alt emakeelekõneleja ja teise keele või võõrkeeleõppija erinevust, mis traditsioonilise lingvistilise analüüsi puhul on enamasti varjatud või tagaplaanile jäänud. Tuginetakse abstraktsioonile, mida nimetatakse grammatikaks, ja ignoreeritakse keeles olevaid loomulikke liigitusi, mis leiavad kajastamist lingvistiliste andmete arvutipõhises ja statistilises analüüsis (vt Abney 1995). Seetõttu on keeleõppe elulähedasemaks muutmise vajadusest tegeldud kõige muu hulgas ka õpikute sisu ning keele analüüsiga. Omal ajal moodustati Tartu Ülikooliski töörühm, kelle ülesanne oli muukeelse kooli jaoks PHARE eesti keele õppe programmi raames koostatud 4.–9. klassi õpikute keerukuse analüüs. Eesmärk oli veenduda, kas õpikud muutuvad aste-astmelt raskemaks. Selleks hinnati õpikute sõnasagedust, sõnade ja lausete pikkust ning võrreldi saadud tulemusi omavahel ja Tartu Ülikooli eesti kirjakeele korpuse vastavate andmetega. Eraldi määrati sõnade abstraktsuse / konkreetsuse aste, koostati õpikutes kasutatud pärisnimede loend, plaanis oli 5000 sõnast koosneva abstraktsuste sõnastiku loomine jne. Nende ülesannete täitmine vajas vastava tarkvara arendamist (Asser *et al* 2004). Tööst kasvas välja eesti kirjakeele sagedussõnastik (Kaalep & Muischnek 2002).

2. Interaktiivsed õppematerjalid ja õpikeskkonnad

See keeleõppe arendamise suund ei ole otseselt olnud seotud korpustega, kuid kahtlemata on nendest abi olnud õppetekstide valikul, adapteerimisel ja harjutuste koostamisel – kõne all on interaktiivsete õppematerjalide ja õpikeskkonna loomine nagu *Virtual Language Centre*[3] või *Eesti e-ülikoolid* ja *e-kutsekoolid*[4]. Haridustehnoloogidel on välja pakkuda erinevatel eesmärkidel

[3] http://www.edict.com.hk/vlc/, 6.07.07

[4] http://www.e-uni.ee/index.php?main=54, 9.09.2007

kasutatavaid e-õppe keskkondi (nt WebCT, Moodle, IVA) ja programme e-kursuste loomiseks (nt Hot Potatoes, Flickr jt). Ka keeleõpetajad on selle suunaga kaasa läinud, nt K.Uibu e-kursus "Akadeemilise teksti loomine" (pälvis konkursil "Aasta e-kursus 2006" eripreemia kodutööde mitmekesisuse ja disaini eest). Õppematerjal sisaldab eritüübilisi ülesandeid, mis K. Uibu sõnul "nõuavad (enese) analüüsi ja individuaalset sooritust, paaris- ja väikestes rühmades tööd ning ühisarutelusid". Õppematerjalide ja tööülesannete koostamisel on ta toetunud Bloomi kognitiivsele taksonoomiale, mille kohaselt "rakendamine pole võimalik ilma mõistmiseta, analüüs ilma rakendamiseta ega süntees ilma analüüsita. Üliõpilased on tõstnud esile kursuse selget ja loogilist struktuuri ning kasutajasõbralikkust"[5].

Interaktiivsete keelematerjalide alla kuuluvad meil ka *Efant*[6] ja *Kaunis külaline*. *Efant* sisaldab kokku 60 Riikliku Eksami- ja Kvalifikatsioonikeskuse lugemis- ja kuulamistesti (kumbagi 30, iga testi juures on keskmiselt 10 harjutust, seega kokku umbes 600 harjutust), sõnastikku ja jututuba. Õpilane saab valida kolme raskusastme vahel: kerge, keskmise ja raske (vt Rummo 2004: 165). *Kaunis külaline* on CD-ROMil töötav eesti keele õppe materjal, mis on loodud erinevate keeletasemete jaoks USAs koostatud õppeprogrammi alusel ja sisaldab lisaks tekstidele ka videomaterjali. Videod leiab kasutaja internetist[7]. Kõige uuem arvutipõhine materjal – „Eesti keel ja meel" (Pangloss 2007) – on audiovisuaalsete ja graafiliste vahendite abil loodud süsteem eesti keele õppimiseks eesti kultuuri kontekstis, mis toimib 8 keele baasil (vene, inglise, saksa, prantsuse, itaalia, kreeka, flaami, ungari ja rumeenia)[8]. Laserplaadil "25 X EESTI" on välja

[5] Vt E-õppe uudiskiri, suvi 2007, lk. 4, http://portaal.e-uni.ee/uudiskiri, 9.09.2007

[6] http://www.efant.ee/student, 6.07.07

[7] http://www.meis.ee/kk/, 6.07.07

[8] http://www.meis.ee/pictures/Artem_Davidjants.pdf, 24.09.2007

antud eesti kõnekeele ja kodakondsuse alane sõnavara, mida õppija saab harjutada (õige tõlkevariandi valimine, lausete koostamine). Algajale ja kesktaseme keeleõppijale on mõeldud CD-ROM "Talk Now!", mis on tehtud Antwerpeni ülikooli mitmekeelse SMALLINC-projekti[9] raames (vt Rammo & Tael 2004: 156). Eesti keele algõpet saab teostada ka arvutiprogrammiga *Oneness on-line language training courses*[10]. Samas keskkonnas võib õppida veel soome, leedu, poola ja portugali keelt. Kursus koosneb 10 ühtse skeemi alusel koostatud õppetsüklist, igas grammatikaseletus, sõnavara osa, fraasimoodustus, harjutused keelekompetentsi arendamiseks ning sotsiokultuuriline info. Test on antud interaktiivse keelemängu vormis, kursusel osalejad saavad omavahel suhelda jututoas.

3. Õppijakeelekorpused

Õppijakeelekorpus on elektrooniline andmekogu, mis koosneb teise keele või võõrkeeleõppija loodud kirjalikest tekstidest ja / või suulise kõne näidetest. Neid koostatakse eesmärgil saada objektiivseid andmeid selle kohta, miks õppija vahekeelt ehk õppijakeelt võib pidada omaette keelevariandiks. Uurijate töölaual on autententse õppijakeele näited, mille alusel saab teha nii teoreetilisi kui pedagoogilisi järeldusi, et seejärel viia vastavusse õppija vajadused ja keeleõpe (vt Granger 1998: 6; Leech 1998).

Teisalt on õppijakeele korpusi kasutatud kommertseesmärkidel, nt inglise sihtkeelega *Cambridge Learner Corpus* ja *Longman Learners´ Corpus* (Granger 2004: 129–130). Esimene neist (CLC) on suunatud keeleteadliku õppija vajadustele. Üle maailma

[9] http://www.isoc.siu.no/isocii.nsf/projectlist/90249, 06.07.07
[10] http://www.oneness.vu.lt/en/, 06.07.07

kokku kogutud keeletestide alusel on tehtud kindlaks, mislaadi vigu üldse esineb. Seda teavet on kasutatud uute õppevahendite, õpikute, sõnastike ja tasemetestide koostamiseks, millele tuginedes oleks võimalik vigu vältida ning samas ka vigade kohta infot saada. Keeletestide analüüsi alusel võib otsustada, missugune tee edasisteks õpinguteks oleks kõige mõttekam, et jõuda oma oskustes järgmisele keeletasemele. Teises kommertskorpuses (LLC) on tähelepanu keskmes spetsiaalsete sõnastike koostamine, mida õppija vajab omandamaks teatud leksika ja grammatika, mis kindlustab talle ka teatud tasemeoskused. Õppija leiab neist sõnastikest just selle sõna, mida tal hetkel diskursuses vaja läheb. Vea tekkimisel suunatakse ta abi järele ja juhatatakse kätte õige tee (vt Pravec 2002: 88–89).

Uurimistöö ja kommertseesmärkidel loodud õppijakeelekorpuste töötlemine ning kasutamine ei ole teineteist välistavad suunad. Ometi on selgeid märke nende liinide lahushoidmisest. Näiteks Louvaini ülikooli *International Corpus of Learner English* (ICLE) on ennast määratlenud teadustööle avatuna ja välistanud andmebaasi kasutamise kommertseesmärkidel, nagu ka Uppsala ülikooli USE (*Uppsala Student English*) korpus, mis on loodud pedagooglistel eesmärkidel (keelevigade raskusastme mõõtmine erinevatel keeletasemetel). Samas aga prantsuse sihtkeelega FRIDA ja *FreeText* valmisid rahvusvahelise Euroopa Liidu projektina just tänu ülikoolide ja ärisektori koostööle (Manchesteri ülikool, Genova ülikool ning prantsuse firma Softissimo). *FreeText* tugineb uuematele teise keele omandamise teooriatele ning on mõeldud kesk- ja kõrgtasemel prantsuse keele õppijate kommunikatiivsete oskuste parandamiseks. Korpusel on automaatne veamärgendussüsteem, töötlemisel on kasutatud erinevaid tarkvarapakette, mis on võimaldanud üle minna **arvutipõhise keeleõppe uuele tasandile – õppijakorpusele**. *FreeText* sisaldab interaktiivset õpikeskkonda, kus on võimalik koostada viit tüüpi kirjalikke tekste (õpetaja saab

neid vastavalt vajadusele kohandada). Samas keskkonnas võib tehtud töö grammatikakontrollijaga üle kontrollida ning saada vajalikke juhiseid elektroonilisest grammatikakäsiraamatust ja sõnastikust. Automaatse veadiagnoosimise süsteemi aluseks on mõjustus- ja sidususteoorial põhinevad tehnikad (vt Vandeventer Faltin 2003). Automaatne vealeidja analüüsib õppijakeele kõiki tasandeid, parandab vead, annab tagasisidet, viidates konkreetsetele grammatikareeglitele. Õppijakorpuste loomine, mis on teoks saanud tänu ülikoolide ja ärisektori koostööle, on andnud tõhusa panuse nii keeleõppe teooria arendamisse kui praktilisse keeleõppesse. Avardunud on arusaamad õppijakeelekorpuste koostamise ja arendamise võimalustest, veadiagnostikast ning keele automaattöötlusvahendite kasutamisest vigade leidmisel ja märgendamisel (vt Eslon & Metslang 2007: 103).

Võrreldes ülalkirjeldatuga on eesti keele õppe võimalused esialgu piiratumad.

Tallinna Ülikooli üld- ja rakenduslingvistika õppetooli projektina on valminud *Eesti vahekeele korpus*[11], mis on loodud uurimistöö ja eesti keele kui teise keele või võõrkeele õppe eesmärkidel. Selle kasutajad on teadustööga tegelevad inimesed ja keeleõpetajad. Lisaks õppijakeele kirjalikele tekstidele (dokumentidele) sisaldab korpus ka metateavet:

1. Teksti koostaja kohta: sugu, vanus, emakeel, kodune keel, haridus, päritolupiirkond Eestis või teistes riikides, sotsiaalne taust – õpilane, üliõpilane, teenistuja, töötu, pensionär jne, teksti alusel määratud keeletase – A, B ja C. Teksti koostaja nime ei ole võimalik tuvastada, kuna töö originaalvariant on hävitatud ning isikuandmeid pole kusagil fikseeritud.

2. Andmeid teksti kohta: maht sõnades ja lausetes, teksti liik (essee, kiri, harjutus jne), teksti koostamise laad (kirjutatud

[11] http://evkk.tlu.ee, 20.09.2007

abivahendeid kasutamata ehk spontaanne tekst / abivahendeid kasutades ehk konstrueeritud tekst). Eelistatud on spontaanseid esseeliiki tekste.

3. Andmeid teksti sisestaja ja märgendajate kohta: avalikustatud on sisestaja eesnimi, märgendajate nimed jäävad avalikkuse eest varjule.

Eesti vahekeele korpus on osaliselt käsitsi märgendatud (ligi 217 000 sõnet, neist ligi 29 000 vigast), tuumkorpuse maht moodustab 500 000 sõnet, korpuse üldmaht seisuga september 2007 oli ligi 608 000 sõnet. Korpus sisaldab vene lähtekeelega õppijate eesti keele kui teise keele terviktekste, mille maht pole rangelt piiratud ning varieerub 50 ja 1000 (harvem enamagi) sõne vahel. Alustatud on allkorpuste loomist, kuhu kuuluvad eesti õppijakeele näited teistest sugulas- ja mittesugulaskeeltest (esmalt soome, saksa ja inglise keelest), mis perspektiivis suurendab korpuse mahtu umbes 1,5 miljoni sõneni ja võimaldab uurida eesti õppijakeelt nii teise keele kui võõrkeele omandamise teooriate võtmes.

Eesti vahekeele korpuse kasutajaliides teeb korpuse interneti kaudu vabalt kättesaadavaks. Korpusel on oma konkordantsileidja, sõna- ning vormisageduse statistika, märgendatud vigu saab näha vealiikide kaupa (leksikaalsed, leksikaalgrammatilised, morfonoloogilised, morfoloogilised, morfosüntaktilised, süntaktilised, kommunikatiivsed) ning kitsamas kontekstis, vajadusel ka terviktekstis. Iga vealiigi all on rohkem või vähem rikkalik alamliigituste hierarhia. Näiteks jagunevad morfosüntaktilised vead järgmiselt: *ma*-infinitiivi kasutamine, *da*-infinitiivi kasutamine, *ma*-infinitiivi käändeliste vormide kasutamine, *des*-vormi kasutamine, kesksõnade kasutamine (pre- ja postpositsioonis), rõhumäärsõnade kasutamine, modaalsõnade kasutamine, järellidete *-gi* ja *-ki* kasutamine. Süntaksivigade jaotus on mitmeastmeline (sõnaühendi süntaks, fraasisüntaks, moodus-

tajate süntaks, lausesüntaks, lauseliikmete ärajätt, ülearune sõna lauses) ning seetõttu keerulisem (nt sõnaühendi süntaksi all on rektsioon ning ühildumine põhisõnaga arvus ja käändes; rektsioon jaguneb verbi- ja noomenirektsiooniks, ühildumise all on ära toodud parataks) jne.

Dokumentide ja andmete esitamiseks on kasutatud XML-formaadi XHTML-versiooni, märgendite hierarhias on tarvitusel XPATH-keel.

Korpuses on loodud eraldi võimalus individuaalseks uurimistööks. Selleks on avatud vastav keskkond. Töö tulemused aitavad arendada olemasolevat vigade taksonoomiat, täpsustada vealiikide hierarhiat ja välja tuua vigade vastastikust sõltuvust ehk vealiikide vahelisi seoseid (nt M. Kitsnik on seda võimalust kasutanud oma magistritöös eituse väljendusvahendite uurimiseks vene lähtekeelega õppijate eesti keele kui teise keele omandamisel – Kitsnik 2007). Sama keskkonda saab kasutada ka korpuspõhise õppetöö korraldamiseks: vigade leidmine terviktekstist, nende diagnoosimine ja defineerimine, põhjuste väljatoomine ja reastamine raskusastme järgi, parandatud variandi kontrollimine jne. Neid ülesandeid on kavas sooritada kahel tasandil – harjutamine koos automaatse tagasisidega ja eksamitöö, kus tagasiside tuleb õpetaja pandud hinde ning kommentaaridega. Nii tehakse *Eesti vahekeele korpuse* baasil algust interaktiivse ja korpuspõhise keeleõppe ühendamisega.

Keeletehnoloogiline arendustöö Eestis on seotud riikliku programmiga "Eesti keele keeletehnoloogiline tugi (2006–2010)"[12]. Üks kavandatavatest rakendustest on süntaksianalüüsil põhineva tarkvara, sh automaatse grammatikakorrektori väljatöötamine. Selleks on vaja suurendada olemasoleva eesti kirjakeele korpuse mahtu ja luua uusi keeleressursse, nt

[12] http://www.hm.ee/index.php?popup=download&id=4964, 11.09.2007

mitmekeelne paralleelkorpus, milles on esindatud eesti, inglise, saksa, prantsuse, soome ja vene keel ning mida läheb vaja automaatsete tõlkeprogrammide töö täpsemaks tegemiseks. Teisalt peetakse oluliseks koostada „vigade korpus", kus paralleelselt grammatiliselt vigase lausega on esitatud sama lause „õige" vaste. Ka see on omamoodi paralleelkorpus, mille peal saab katsetada grammatikakorrektorit ja luua keeleõppeprogramme. Esialgu on olemas 50 000 sõnest koosnev korpus, mille mahtu kavatsetakse suurendada 200 000ni, seejärel tehakse korpus Internetis kättesaadavaks. Kaugem eesmärk on nendele ressurssidele üles ehitada veebipõhine interaktiivne eesti keele õpe. Eesmärgini jõudmiseks peab grammatikakorrektor suutma teksti morfoloogiliselt ühestada ning kontrollida rektsiooni- ja ühildumisvigu. Iga keeleteadlase ja keeletehnoloogi unistus on, et ka eesti keele alusel oleks võimalik analüüsida ja sünteesida ebastandardset teksti. Selleks peab grammatiline analüüs ja süntees muutuma põhjalikumaks ning täpsemaks, mis omakorda eeldab rea keeleteaduslike probleemide lahendamist (nt ühend- ja väljendverbide ning kollokatsiooniliste sõnaühendite tuvastamine, vt Roosmaa *et al* 2001; Muischnek *et al* 2003: 69–75; Muischnek 2006; Pool 2007: 44–66). Kirjeldatud suundi arendatakse põhiliselt Tartu ülikoolis.

Õppijakeele- ja õppijakorpuste vajalikkust on tunnetanud ka soome teadlased, kes on seni keeleõppe eesmärgil edukalt paralleel- ning tõlkekorpuste võimalusi kasutanud (Joensuu ja Jyväskylä Ülikool, Tampere Ülikooli vene-soome paralleelkorpus – vt eespool *paralleel- ja tõlkekorpuste kontrastiivne ning tõlkeanalüüs*) või koostanud eksamitekstikogusid nagu Helsingi Ülikooli soome keele ja kirjanduse osakonna "muud materjalid", kuhu on koondatud soome kui emakeele, teise keele ja võõrkeele riigieksamikirjandid[13]. Nüüdseks on Oulu Ülikoolis otsus-

[13] http://www.helsinki.fi/hum/skl/tutkimus/suomi.htm, 27.08.2007

tatud luua esimene soome õppijakeele korpus (J. Jantunen, H. Sulkala). Töö esimesel etapil on korpusesse kavas koondada vene, rootsi ja eesti emakeelega õppijate keelenäited. Eesti-poolne partner on Tallinna Ülikooli eesti keele ja soome keele (võõrkeeltena) õppetool (A. Kaivapalu), koostööd tehakse samuti *Eesti vahekeele korpuse* töörühmaga. Soome õppijakeele pilootkorpuse andmeanalüüsi tulemusi on tutvustanud J. Jantunen (2007a; 2007b).

Toodud ülevaade erinevate korpuste kasutamisvõimalustest keeleõppes viitab sellele, kuivõrd arenenud on ala maailmas (eraldi võttes inglise, rootsi ja prantsuse keel) ning missugune tee on käia eesti keele kui teise või võõrkeele õpetamiseks vajalike adekvaatsete õppematerjalide, sõnastike, õpikute, pedagoogilise grammatika koostamisel, korpuspõhise keeleõppe väljaarendamisel, rakendamisel ning vastaval uurimistööl. Keeletehnoloogiliseks ja lingvistiliseks sõlmküsimuseks jääb seejuures automaatne vealeidja.

Mis iseloomustab *Eesti vahekeele korpuse* veaanalüüsi hetkeseisu?

Eesti vahekeele korpuse veaanalüüs tugineb mitmemõõtelisele lingvistilisele taksonoomiale, mida käesoleva artikli autor on varem juba tutvustanud (Eslon 2006b: 14–17; Eslon & Metslang 2007: 106–112). Seetõttu antakse siinkohal ülevaade korpuse tekstide märgendamisest ning sõna- ja vormisageduse seisust.

Märgendamise (*tagging*) all mõistetakse *Eesti vahekeele korpuses* veamärgendust, mille tulemusena eraldatakse üksteisest korrektne ja grammatikanormidest kõrvale kaldunud õppijakeel. Nende suhet teineteisesse saab uurida sõnakaupa või vealiigiti kas ühes tekstiliigis või korpuses tervikuna, samuti lause, teksti-

lõigu või tervikteksti tasandil. Saadud lingvistilist teavet on võimalik siduda metalingvistilise info ning sõna- ja vormistatistikaga

Märgendamine toimub käsitsi, esialgu puudub Eestis ühtne teadmiste pank ja töökeskkond, mis ühendaks keeleressursid ning tarkvara (vt Viks 2002). Samuti pole elujõuliseks saanud grammatikakorrektor (vt ülal). Märgendussüsteem on mitmetasandiline: iga ülemtasandi märgend sisaldab endas selle kõikide alamtasandite märgendeid. Kui märgendaja võtab töösse uue dokumendi, siis avanevad talle esimesena vigade ülemliigid (Leksikaalsed, Leksikaalgrammatilised, Morfonoloogilised, Morfoloogilised, Morfosüntaktilised, Süntaktilised, Kommunikatiivsed), millele on lisatud Proovi kätt ja Sõnatuletus. Ülemliik "Proovi kätt" on mõeldud juhtumite jaoks, mida pole võimalik kokku viia veataksonoomia ühegi liigi või alamliigiga. Klikkides kindlale vealiigile avanevad selle liigi alamliigid, klikkides ühele alamliikidest – selle alamliigid jne. Vt näide:

○ Morfonoloogilised
    ○○ Astmevaheldus
        ○○○ seoses sõnatuletusega
        ○○○ seoses vormimoodustusega
        ○○○ tüvevaheldus ja supletiivsed tüved
        ○○○ deminutiivsete liidete kasutamine

Märgendaja valib klassifikatsiooni ülemliikide hulgast vajaliku alamliigi ja ühendab märgendi vastava vigase kohaga õppijatekstis. Vead kuvatakse tekstis, vealiigid aga rippmenüüs. Teksti vigane osa piiritletakse vasakult ja paremalt: >> ja << . Hõlpsamaks jälgimiseks on korpuse disainis sel eesmärgil kasutatud punast värvi. Näiteks:

Minu unistuste auto >>.<<

Minu unistuste auto peab olema >>moodsus<< , >>kiirus<< , ilus… Sellepärast mulle ei meeldi >>vanad autod ja liiga >>väiked<<<< . Tahan, et minu auto oleks >>mugavus<< ja >>pehmed istmed<< . >>Auto peab olema taskukohane<<, et ma saaksin >>osta<<. Ei taha, et oleks raske >>juhimine<< >>,<< suur kütusekulu.

Märgendatud teksti parempoolsest rippmenüüst leiab korpuse kasutaja metainfo teksti ja selle kirjutanud inimese kohta, mille all on tekstile lisatud märgendite nimistu. Kui viia kursor ühele vealiigile nimistust, siis kuvatakse see viga tekstis. Mugavamaks jälgimiseks on artiklis kasutatud numbreid (1), (2) ... (11):

Minu unistuste auto (1) >>.<<

Minu unistuste auto peab olema (2) >>moodsus<< , (3) >>kiirus<< , ilus… Sellepärast mulle ei meeldi (4) >>vanad autod ja liiga väiked (5) >>väiked<<<< . Tahan, et minu auto oleks (6) >>mugavus<< ja (7) >>pehmed istmed<< . (8) >>Auto peab olema taskukohane<< , et ma saaksin (9) >>osta<< . Ei taha, et oleks raske (10) >>juhimine<< (11) >>,<< suur kütusekulu.

Tekstis märgendatud vealiigid.

(1) Interpunktuatsioonivead
(2) Vale sõnaliigi kasutamine
(3) Vale sõnaliigi kasutamine
(4) Sõnajärg ja lause teatestruktuur
(5) Omadussõna käändevormide moodustamine ja kasutamine
(6) Vale sõnaliigi kasutamine
(7) Verbirektsioon
(8) Semantiline seos sõnade vahel
(9) Tegevuse transitiivsus / intransitiivsus
(10) Vale sõnaliigi kasutamine
(11) Sidendite kasutamine olenevalt seose semantikast

Metainfost võib teksti autori kohta välja lugeda, et ta on Ida-Virumaalt pärit vene emakeelega naine, vanusepiiriga 26–40 aastat, kuulub teenistujate hulka, omab keskharidust ja valdab eesti keelt A-tasemel. Oma töö on ta kirjutanud spontaanselt, st abivahendeid kasutamata ja seega klassitunnis. Teksti kohta saame teada, et tegu on vastusega küsimusele, mis sisaldab viit lauset ja 47 sõna. Märgendaja on leidnud 11 viga, neist üks esineb korduvalt – märgend "vale sõnaliik" oli lisatud neljal korral, vt (2), (3), (6), (10). Järelikult on märgendatud õppijakeele tekstis leitud kaheksa erinevat vealiiki. Näiteks:

Informant	**Tekst**
Sugu: naine	Tüüp: vastkys
Vanus: kuni 40a	Sõnu: 47
Elukoht: Ida-Virumaa	Lauseid: 5
Sots.: teenistuja	Vigu kokku: 11
Emakeel: vene	Erinevaid: 8
Kodus: vene	
Haridus: kesk	
Tase: A	
Abivahendid: ei	

Edaspidine töö keelevigade analüüsimisel on seotud nende jaotumisega erinevate veaklasside vahel. Veaklassid on abstraktsioon, mis tugineb arusaamale keelesüsteemi paradigmaatilisest ja süntagmaatilisest ülesehitusest ning mis kujuneb nende kahe telje ristumispunktides. Paradigmaatilisteks tunnusteks on hierarhia grafeem → morfeem → sõna → sõnaühend → lause → tekst, mis süntagmaatilisel teljel on seostatud keelesüsteemi kolme aspektiga – semantika, grammatika ja pragmaatika. Kokku tekib 18 veaklassi, kui teksti tasand välja jätta, siis 15. Näiteks:

1 – grafeem + semantika (*need inimesed on *laiad / lai**s**ad*, grafeem ***s*** eristab sõnu)

2 – grafeem + grammatika (*maja tagasi on *õu**e** / õu*, grafeem ***e*** eristab käändevorme)

3 – grafeem + pragmaatika (**K-Järvelt* / *Kohtla-Järvelt*, väljendustava vastu eksimine)

4 – morfeem + semantika (**nad andsid mulle* ***tarku*** *selles õppeaines* / *tarku**st** selles õppeaines*, morfeemi ärajätt on sõnu eristav tunnus)

5 – morfeem + grammatika (*Sa *oska**si**d palju huvitavaid faktid ajaloost* / *sa* ***tead*** *palju huvitavaid fakte ajaloost*, mineviku ajavormi kasutamine ei sobi kokku edastatava informatsiooniga)

6 – morfeem + pragmaatika (****vot*** *meie valitsuses* / *meie valitsuses*, venepärane morfeem teksti sidususfunktsioonis)

7 – sõna + semantika (**inimesed peavad tihti* ***mõlgutama*** *meie keskkonnast* / *peavad* ***mõtlema*** *meie keskkonnast*, eksimine sõnade semantilise seose vastu)

8 – sõna + grammatika (**aadressiltänav* / *aadress*, vale liitsõnamoodustusmall)

9 – sõna + pragmaatika (**valdab* ***ainult*** *teadmisega* / *valdab teadmisi*, ebasobiv rõhutamine) jne.

Ühe veaklassi piires ning veaklasside vahel moodustub vealiikide ja alamliikide vaheliste seoste võrgustik. Eeldatavalt võivad seosed olla korrelatiivsed, mitmesuunalised, üksteist ja / või teineteist välistavad, hajusad. Nende väljaselgitamisel õnnestub ilmselt jälgida, missugune viga millist saadab, missugune millisest tuleneb, et reastada vead raskusastme järgi. Saadav teave on oluline empiiriline materjal pedagoogilise grammatika koostamiseks, kuna kirjeldus tugineb reaalse õppijakeele veaanalüüsile. Oma täpsustused keelevigade raskusastmete mõõtmisse lisab vigaste vormide kasutussageduse võrdlus sama sõna õigete vormide esinemisega õppijakeeles.

Eesti vahekeele korpuse praegusel arendusetapil on sõna- ja vormisageduse statistika kättesaadav korpuse tekstide sõnavara ulatuses (*count*) või algustähe järgi (*word*). Esimest loendit annab võrrelda eesti kirjakeele sagedussõnastiku tuhande sagedama sõnavormi tabeli andmetega (Kaalep & Muischnek 2002: 174–186) ja teist – tuhande sagedama sõnavormi tähestikulise järjekorraga (Kaalep & Muischnek 2002: 187–199). Andmete võrdlemisel tuleb silmas pidada, et eesti kirjakeele korpuse valim on õppijakeelekorpuse valimist poole suurem (vastavalt 1 miljon ja 500000). Siinkohal esitame kolmekümne sagedasema sõnavormi võrdluse:

Sagedus	**Eesti õppijakeel**	**Eesti kirjakeel**	Sagedus
13939	Ja	Ja	27214
13295	On	On	19184
5553	Et	Ei	13810
5091	Ei	Et	12314
5087	Ma	Ta	10170
4073	Oli	Oli	8861
3882	Eesti	Kui	8599
3835	See	Ka	6191
3556	Kui	See	6114
2991	Ka	Oma	5329
2478	Oma	Aga	5274
2340	Aga	Ma	4454
2337	Ta	Ning	4409
2230	Väga	Mis	4391
1933	Mis	Siis	4238
1856	Minu	Nii	3523
1741	Palju	Või	3310
1376	Kõik	Seda	3177
1290	Nad	Kes	2961
1260	Sest	Nagu	2961
1246	Või	Tema	2720
1243	Nii	Veel	2550 →

1233	Siis	Pole	2534
1197	Me	Kuid	2341
1189	Olid	Selle	2337
1182	Seda	Kas	2287
1147	Mida	Juba	2234
1073	Mulle	Nad	2190
1058	Selle	Välja	2008
1041	Ning	Midagi	1978

Võrdlusnäite alusel saab teha kõige üldisema järelduse, et õppija- ja kirjakeele vahel on selgeid ühisjooni: sagedasemateks sõnaklassideks on side- ja asesõnad, *olema*-verbi oleviku ja lihtmineviku vormid, rõhusõna *ka.* Nimetatud sõnaklassid on olulised tekstiloomes, kuna neid kasutatakse laialt teksti sidususvahendite funktsioonis, samuti deiktilise nullpunkti tähistamiseks (mina – siin – praegu). Ajadeiktilistest sõnadest on kolmekümne sagedasema hulgas *siis*, mida kirjakeeles on veidi rohkem kasutatud. Tuumverbi *olema* vormid vastandavad olevikku ja minevikku. See on näide õppijakeele lingvistilistest universaalidest. Teine silmahakkav moment on seotud isikuliste asesõnade *ma* ja *ta* korrelatiivsusega sagedusreas. Siin on tegu kahe keelevariandi lahknevusega: õppijakeeles valitseb *ma*-alge ja kirjakeeles *ta*-alge. Tasub uurida, kas erinevus võib olla seotud pragmaatiliste asjaoludega – vene keelele on omane rõhutatult *mina*-keskne esituslaad, mis on avaldanud mõju ka vene lähtekeelega õppija eesti sihtkeelele. Kolmas nähtus, mis statistikat vaadates silma jääb, on asesõnade üldine rohkus esimese kolmekümne sagedasema sõna hulgas ning nende sageduse teatud korrelatiivsus õppija- ja kirjakeeles (vastavalt *ma – ta; oma – oma; ta – ma; minu – Ø; nad, me – tema; mulle – nad*) jne. Õppijakeele eripära võib seostada ilmselt sõnade *Eesti* ja *palju* esinemisega kolmekümne sagedasema hulgas. Kirjakeeles seda tendentsi ei täheldata. Tegu võib olla jällegi prag-

maatilist laadi nähtusega nagu keeleõppe materjalide temaatiline jaotus, mis tõstab esikohale Eestiga seotud teabe (nt "Eesti – meie ühine kodumaa"). Määrsõna *palju* ülekasutus võrreldes kirjakeelega on ilmselt nähtus, mis kuulub emakeele interferentsi alla. Miskipärast valivad vene emakeelega õppijad eesti vahekeeles just selle määrsõna, et midagi rõhutada või esile tuua. Siinkohal on tegu selge lahknevusega emakeele- ja eesti keele kui teise keele kasutaja tekstiloomes.

Sõnavormide tähestikuline järjekord võimaldab saada veidi teistlaadset teavet. Võtame näiteks ***a***-tähe alguses olevate nimisõnade ja nende tuletiste vormid (tärniga on välja toodud vigased, number näitab sagedust korpuse tekstides):

> aabikaasa 1, *aabikasa 1, aabits 7, aabitsaid 3, aabitsast 3, aadli 5, aadliga 2, *aadlike 6, aadlikke 3, aadliku 1, aadlikud 6, aadlil 2, aadliperekonna 2, *aadlis 2, aadliseisusest 1, aadlisoost 3, aadlivastane 2,*aadres 3, *aadresi 1, aadress 13, aadressi 2, aadressil 1, *aadressiltänav 1, aadressiraamat 2, aadressiraamatud 1, *aadrissiks 1 jne.

Hetkel pole seda loendit veel võimalik üksüheselt võrrelda kirjakeele 1000 sagedasema sõnavormiga tähestikulises järjekorras, kuna õppijakeele 1000 sagedasemat ei ole veel eraldi välja toodud – need tuleb loendist käsitsi välja noppida. Seetõttu algabki kirjakeele loend nimisõna *aasta* vormidega *aasta* 1130, *aastaid* 98, *aastal* 779, *aastat* 642, *aastate* 117 ja *Eesti vahekeele korpuse* statistikas sõnaga *abikaasa* jt. Nimisõna *aasta* leiab asjasthuvitatu sealt, kus see sõna tähestikulises järjekorras olema peab. Esindatud on järgmised vormid ja tuletised:

> aasta 269, aastaaeg 4, aastaajast 1, aastaarv 1, aastaarve 3, *aastaas 1, aastad 28, *aastade 1, aastaga 8, aastaid 15, aastail 5, *aastak 1, aastaks 19, *aastakssee 1, aastakäikude 1, aastakümned 1, aastakümneid 5, aastakümnetel 3, aastal 486, aastale 3, aastalt 25, aastana 1, aastane 56, aastani 19, aastapäeva 1, aastapäeval 1, aastapäevale 4, aastas 44, aastasadade 1, aastasajaks 2, aastased 2, aas-

taseks 2, aastaselt 1, aastasena 3, aastast 51, aastastele 4, aastat 327, aastate 36, aastatega 2, aastatel 71, aastateni 2, aastatesse 2, aastatest 1, *aastatkõrgkoolide 1, *aastattel 1, *aastatuh 3, aastatuhande 6, aastatuhandeid 3, aastatuhandel 3, aastatuhandete 2, aastatuhat 3, aastavahetus 2, aastavahetuseks 1, aastavahetusel 1, aastavahetuseprogrammid 1, *aasttuh 3, *aastunud 1, *aatat 1.

Kui sellest loendist välja noppida vaid kõige sagedamini kasutatud vormid (100 ja enam korda), siis saame järgmise pildi: *aasta* 269, *aastal* 486, *aastat* 327. Mitmuse osastava (*aastaid* 15) ja omastava käände (*aastate* 36) vorme oli kasutatud kirjakeelest tunduvalt vähem. Statististiliste andmete võrdlus näitab, et kirjakeele sagedased sõnavormid on õppijakeeles küll korrektsed, kuid kahte neist on ilmselgelt alakasutatud. Põhjuste selgitamine vajab täpsemat analüüsi. Siinkohal piirduks autor ühe huvipakkuva faktiga: eesti keele formaalse grammatikaga seoses on uurijad märkinud, et morfoloogilise ühestamise käigus tegi ühestaja tekstikorpuses 180 viga (kustutas 1,9% sõnade õige tõlgenduse), mis oli seotud enamasti nimetava, osastava ja lühikese sisseütleva eristamisega (Roosmaa *et al* 2001: 92). Ilmselt võib ühe põhjendusena viidata sellele, et morfoanalüsaatori töökindlus eesti keele kirjeldamisel sõltub keelesisestest raskustest – olemasolev deskriptiivne grammatika ja reaalne keelekasutus ei ole vastavuses. Eesti keelt emakeelena kõnelevate inimeste suulises keelepruugis, samuti veebikeeles on saanud harjumuspäraseks eksida sihitise käändevormi valikus (nt *Eile valiti presidenti*; *Nad on välistanud seda, et ...*). Reaalsus on, et tegevuse piiritlematuse / piiritletuse markeerija – sihitise kääne – on tasapisi hakkanud oma funktsioone kaotama. Emakeelekõneleja ei näe siin probleemi, keeleveana hindavad seda vaid lingvistid. Järelikult pole imekspandav, et eesti õppijakeeles püütakse neid vorme vältida – õpikureeglite ja tegeliku keelekasutuse vahel puudub selgepiiriline vastavus.

Edasine töö *Eesti vahekeele korpuse* veamärgenduse arendamisel on seotud morfoloogilise analüüsiga. Eesti keeletehnoloogilistest rakendustest tundub selleks sobivaim olevat Ü. Viksi loodud Eesti Keele Insituudi süsteem, mis võimaldab koostada õppijakeele korrektse keelekasutuse elektronsõnastiku, kus on näha tüvemuutused, sõnavormid, grammatiline homonüümia (Viks 2000 : 35). Lisaks saab õigeid tõlgendusi ka nendele sõnadele, mida Ü. Viksi „Väikese vormisõnastiku" elektroonilises variandis ei ole (vt Viks 1994 : 161). Eesti Keele Instituudi analüsaatorit on võimalik kasutada esmajoones uurimistöö eesmärgil. Teine valik on ESTMORFi[14] rakendamine lingvisti töövahendina, milles tundmatute sõnade tuvastamisel saab kasutada oletajat. Analüsaator on reeglipõhine, selle abil on võimalik kindlaks teha, kas sõna on sõnastikus või mitte (vt Kaalep 1998: 23), et seejärel analüüsi etapiviisiliselt jätkata (Kaalep 1996: 73–82). ESTMORF on olnud kommertsrakenduste aluseks (nt lemmatiseerija, õigekirjakontrollija, poolitaja). Ees seisab hulk tööd, katsetusi ja otsinguid, et olemasolevaid keeletehnoloogilisi rakendusi „kombineerides ja täiendades" (Viks 2002) jõuda õppijakeelekorpuse tekstide (pool)automaatse märgenduseni.

Kokkuvõtteks

Suured andmekogud ning keeletehnoloogiliste rakenduste loomine teaduslikel ja kommertseesmärkidel on teinud teise keele / võõrkeeleõppe korpuspõhiseks. Oma koht selles süsteemis kuulub ka paari aasta eest Internetis avatud *Eesti vahekeele korpusele*. Tänaseks on korpuse kasutaja käsutuses konkordantsileidja, vigade otsingut saab teha üle 200000 käsitsi märgendatud sõne

[14] http://www.filosoft.ee/html_morf_et , 15.09.2007

põhjal. Märgendamise aluseks on olnud mitmemõõteline lingvistiline veataksonoomia. Loodud on korpuse sõna- ja vormisagedusmoodul. Omavahel on võimalik ühendada tekste, sõnu, vigu ja õppija kohta käivat metainfot. Alustatud on eesti keele morfoanalüsaatori katsetamise ja õppijakeele sõnastiku koostamisega. Paralleelselt käib töö arvutipõhise keeleõppe loomisel.

Seni tehtu olnuks võimatu, kui puuduksid õppijakeelekorpuse ideedest nakatunud inimesed. Siinkohal tahab autor avaldada siirast tänu õppetekstide kogumisel abiks olnud Signe Abelile, Anne Krivanile, Olga Elksninile, Inguna Joandile, Maarja Heinale ja Elle Väljale Ida-Virumaalt ning Ingrid Krallile, Marju Kõivupuule, Lilian Vanemile ja teistele headele kolleegidele Tallinna ülikooli eesti filoloogia osakonnast.

Kirjandus

Abney, Steven 1996. Statistical Methods and Linguistics. – The Balancing Act: combining symbolic and statistical approaches to language / Ed. by J. L. Klavans, P. Resnik. Cambridge: MIT Press, http://www.vinartus.net/spa/95c.pdf, 8.07.2007.

Asser, Hiie & Kaalep, Heiki-Jaan & Linnas, Siret & Mikk, Jaan & Muischnek, Kadri & Songe, Merje & Uibo, Heli 2004. Õpikute keerukuse analüüs arvutitel. – Toimiv keel II. Töid rakenduslingvistika alalt / Toim. M.-M.Sepper, J.Lepasaar. TPÜ eesti filoloogia osakonna toimetised 3. Tallinn: TPÜ kirjastus, 72 – 84.

Bergh, Gunnar 2005. Min(d)ing English language data on the Web: What can Google tell us? – ICAME Journal. Computers in English Linguistics 29, 25 – 46, http://icame.uib.no/ij29-page25-46.pdf, 14.07.2007.

Corder, Stephen Pit 1981. Error Analysis and Interlanguage. London: Oxford University Press.

Eslon, Pille & Metslang, Helena 2007. Õppijakeel ja eesti vahekeele korpus. – Eesti Rakenduslingvistika Ühingu aastaraamat 3 / Toim. H. Metslang, M. Langemets, M.-M. Sepper. Tallinn: Eesti Keele Sihtasutus, 99–116.

Eslon, Pille 2006a. Analoogiast keelte kõrvutamisel. – Keel ja Kirjandus 1, 15–24.

Eslon, Pille 2006b. Eesti vahekeele korpusest korrelatsioonigrammatikani. – Eesti Rakenduslingvistika Ühingu aastaraamat 2 / Toim. H. Metslang, M. Langemets. Tallinn: Eesti Keele Sihtasutus, 11–24.

Granger, Sylviane 1998. The computer learner corpus: A versatile new source of data for SLA research. – Learner English on computer / Ed. by S. Granger. London: Longman, 3–18.

Granger, Sylviane 2003. Error-tagged learner corpora and CALL: a promising synergy. – CALICO Journal 20(3), 465–480, http://selene.lib.jyu.fi:8080/julpu/9513915425.pdf, 19.09.2007.

Granger, Sylviane 2004. Computer Learner Corpus Research: Current Status and Future Prospects. – Applied Corpus Linguistics. A Multidimensional Perspective / Ed. by U. Connor, T. A. Upton. Amsterdam / New York: Rodopi, 123–145.

Jantunen, Jarmo Harri 2007a. Corpus-driven Analysis Learner Finnish. – 8th Conference on Nordic languages as second languages. Helsinki 10 – 12.5.2007. Abstracts, 24, http://www.helsinki.fi/hum/skl/nordand2007/abstracten.pdf, 9.09.2007.

Jantunen, Jarmo Harri 2007b. Oppijansuomen piirteitä korpusvetoisesti. (*Ilmumas*.)

Jantunen, Jarmo Harri & Eskola, Sari 2002. Käänössuomi kielivarianttina: syntaktisia ja leksikaalisia erityispiirteitä. – Virittäjä 2, 184–207.

Johnson, Keith 2006. Quantitative Methods in Linguistics, http://linguistics.berkeley.edu/~kjohnson/quantitative/, 26.09.2007.

Kaalep, Heiki-Jaan & Muischnek, Kadri 2002. Eesti kirjakeele sagedussõnastik. Tartu: TÜ Kirjastus.

Kaalep, Heiki-Jaan 1998. Tekstikorpuse abil loodud eesti keele morfoloogiaanalüsaator. – Keel ja Kirjandus 1, 22–29.

Kaalep, Heiki-Jaan 1996. ESTMORF: A Morphological Analyzer for Estonian. – Estonian in the Changing World / Ed. by H. Õim. Tartu: TÜ Kirjastus, 43–98.

Kaivapalu, Annekatrin 2006. Reeglid ja analoogia võõrkeeleõppes soome mitmusevormide käänamise näitel. – Eesti Rakenduslingvistika aastaraamat 2 / Toim. H.Metslang, M. Langemets. Tallinn: Eesti Keele Sihtasutus, 71–92.

Kerge, Krista 2007. Euroopa keeleõppe ühtne raamistik ja emakeel. – Oma Keel 14, 27–40.

Kerge, Krista 2006. Eestlase viisakusruumi kajastusi tekstikorpuses. – Tartu Ülkooli eesti keele õppetooli toimetised 30, 67–89.

Kerge, Krista 2003. Aja- ja ilukirjanduse süntaktilise keerukuse dünaamika XX sajandil. – TPÜ eesti filoloogia osakonna veebitoimetised (1), http://www.tlu.ee/fil/veebitoimetised/pdf/lingvistika1.pdf, 4.07.2007.

Kilgi, Annika 2005. Mida eestlane teeb? Verbisageduse muutumisest ainsuse kolmanda pöörde näitel. – Muutuva keele päev 4.11.2005. Teesid, 3–4, http://www.fl.ut.ee/orb.aw/class=

file/action=preview/id=127495/muutuvkeel_teesid.pdf, 4.07.2007. *Vt ka*: Kilgi, Annika 2006. Mida eestlane teeb? – Oma Keel 12, 20–24.

Kitsnik, Mare 2007. Õppijakeele uurimise ja arendamise võimalusi eesti vahekeele korpuse põhjal (erituse väljendamise näitel). Magistrtöö. Tallinn (käsikirja saab lugeda EVKK koduleheküljel, http://evkk.tlu.ee, 13.09.2007).

Kitsnik, Mare 2006. Keelekorpused ja võõrkeeleõpe. – Eesti Rakenduslingvistika aastaraamat 2 / Toim. H. Metslang, M. Langemets. Tallinn: Eesti Keele Sihtasutus, 93–107.

Leech, Geoffrey (1998). Preface. – Learner English on computer / Ed. by S. Granger. London: Longman, xiv–xx.

Mauranen, Anna & Kujamäki, Pekka (Eds) 2004. Translation Universals: Do They Exist? Amsterdam: Benjamins.

McEnery, Tony & Wilson, Andrew 2001. Corpus linguistics / 2nd Ed. Edinburgh: Edinburgh University Press.

Metslang, Helle 2005. Tähelepanekuid verbisüntaksist. – Muutuva keele päev 4.11.2005. Teesid, 7–8, http://www.fl.ut.ee/orb.aw/class=file/action=preview/id=127495/muutuvkeel_teesid.pdf, 4.07.2007.

Metslang, Helle 1994. Temporal Relations in the Predicate and the Grammatical System of Estonian and Finnish. – Oulun yliopiston suomen ja saamen kielen laitoksen tutkimusraportteja 39. Oulu, 216–218.

Михайлов, Михаил 2003. Параллельные корпуса художественных текстов: принципы составления и возможности применения в лингвистических переводческих исследованиях (на примере русско-финского параллельного корпуса художественных текстов). – Acta Universitatis Tampe-

rensis 956. Tamepere University Press, http://acta.uta.fi/pdf/951-44-5754-4.pdf, 9.09.2007.

Muischnek, Kadri 2006. Verbi ja noomeni püsiühendid eesti keeles. – Dissertationes Philologiae Estonicae Universitatis Tartuensis 17. Tartu: TÜ Kirjastus.

Muischnek *et al* = Muischnek, Kadri; Orav, Heili; Kaalep, Heiki-Jaan; Õim, Haldur 2003. Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara / Toim. U. Talvik. Tallinn: Eesti Keele Sihtasutus, http://www.hm.ee/index.php?popup=download&id=3993, 16.09.2007.

Myles, Florence 2005. Interlanguage corpora and second language acquisition research. – Second Language Research 21, 4, 373–391.

Nikolajev, Alexander 2007. Soumen nominaalisen taivutusjärjestelman kvantitatiivista analyysia. – Kielitieteen päivät. Oulu 24.–25. toukokuuta 2007. Abstraktikirja, 75, http://www.oulu.fi/kielitieteenpaivat2007/Abstraktit.pdf, 4.07.2007.

Oakes, Michael P. 1998. Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press.

Roosmaa *et al* = Roosmaa, Tiit; Koit, Mare; Muischnek, Kadri; Müürisep, Kaili; Puolakainen, Tiina; Uibo, Heli 2001. Eesti keele formaalne grammatika. Tartu Ülikooli arvutiteaduse instituut. Tartu: TÜ Kirjastus.

Pool, Liisi 2007. Verbikesksete püsiühendite automaattöötlus. Magistritöö / Juh. K. Muischnek. Tartu: Tartu Ülikooli eesti ja üldkeeleteaduse instituut.

Pravec, Norma A. 2002. Survey of learner Corpora. – ICAME Journal № 26, pp. 81–114, http://icame.uib.no/ij26/pravec.pdf, 11.09.2007.

Rammo, Sirje & Tael, Maarika 2004. Eesti keele õppematerjalid CD-ROMil. – Emakeel ja teised keeled IV. Tartu Ülikooli eesti keele (võõrkeelena) õppetooli toimetised 3. Tartu: TÜ Kirjastus, 156–163.

Rummo, Ingrid 2004. Efant.ee – interaktiivne keeleõppekeskkond vene koolilastele. – Emakeel ja teised keeled IV. Tartu Ülikooli eesti keele (võõrkeelena) õppetooli toimetised 3. Tartu: TÜ Kirjastus, 164–173.

Selinker, Larry 1992. Interlanguage. – Error Analysis. Perspectives on Second Language Acquisition / Ed. by J. C. Richards. London: Longman, 31–54. (Artikli esmatrükk 1972. a.)

Uibo, Heli 2006. Optimizing the finite-state description of Estonian morphology. – Proceedings of the 15th NODALIDA conference, Joensuu 2005 / Ed. by S.Werner. Ling@JoY 1, 203–209.

Uibo, Heli & Bick, Eckhard 2005. Treebank-based research and e-learning of Estonian syntax. – Proceedings of Second Baltic Conference on Human Language Technologies: Second Baltic Conference on Human Language Technologies; Tallinn, Estonia; April 4–5, 2005 / Ed. by M. Langemets, P. Penjam. Tallinn: Institute of Cybernetics, 195–200.

Vandeventer Faltin, Anne 2003. Syntactic Error Diagnosis in the context of Computer Assisted Language Learning. PhD Theses. Genova University, http://www.unige.ch/cyberdocuments/theses2003/VandeventerA/these.pdf, 07.09.2006.

Viks, Ülle 2002. Mis kasu on keeleteadusel keeletehnoloogiast? – Arvutimaailm, http://www.ria.ee/lib/am-2001-2005/3849_5EF.HTM, 16.09.2007.

Viks, Ülle 2000. Eesti keele avatud morfoloogiamudel. – Arvutilingvistikalt inimesele / Toim. T. Hennoste. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: TÜ Kirjastus, 9–36.

Viks, Ülle 1994. Eesti keele morfoloogiline analüsaator. Automaatanalüüsi võimalused ja võimatused. – Keel ja Kirjandus 3, 150 – 163.

Volk, Martin 2002. Using the Web as Corpus for Lingistic Research. – Tähendusepüüdja / Toim. R. Pajusalu, T. Hennoste. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3. Tartu: TÜ Kirjastus, 355–369.

Learner corpora and language learning

Summary

This article deals with current possibilities and needs of language learning. The focus is on learner corpora: development of learner corpora on scientific and commercial purposes as well as the use of them for advancement of foreign language learning and teaching. There have been created language learning systems that enable computer-aided learning, by utilizing e.g. electronic grammars and lexicons. The article gives an overview of the opportunities provided by the Estonian interlanguage corpus. Error tagging and word frequency are viewed more thoroughly.

Keywords: corpora and language learning; Estonian interlanguage corpus; error tagging, taxonomy, and statistics.