ÕPPIJAKEEL JA EESTI VAHEKEELE KORPUS

Pille Eslon, Helena Metslang

Ülevaade. Eesti keele õpetamine võõrkeelena on kutsunud esile vajaduse minna üle käsitsi tehtud väikeuuringutelt õppija vahekeele korpusepõhisele ja osaliselt automatiseeritud uurimisele, kuna suurte andmehulkade analüüs annab objektiivsema pildi eri tasemel õppijate arengust ja vajadustest teel sihtkeele omandamisele. Kirjaliku vahekeele korpuse alusel saab uurida ning arendada kirjutamis- ja lugemisoskust; korpusuuringute tulemusi võib kasutada uute ainekavade, õppevahendite, koolisõnastike ja käsiraamatute koostamisel. Käesolev artikkel tutvustab Tallinna Ülikooli üld- ja rakenduslingvistika õppetooli juures alustatud vene emakeelega õppijate eesti vahekeele korpuse loomise tööd eesmärgiga toetada teise keele omandamise (ingl second language acquisition) teaduslikku uurimist ja eesti keele kui teise keele õpetamise edasist arendamist.*

Võtmesõnad: korpuslingvistika, vahekeel, õppijakeele korpus, veaanalüüs, pedagoogiline grammatika

Sissejuhatavalt

Tänapäeva keeleuuringuid eristab varasematest võimalus teha järeldusi suurte andmekogude alusel, milleks kasutatakse erinevat tüüpi elektrooniliselt töödeldavaid korpuseid. Sama suund on iseloomulik ka võõrkeele omandamise teooria ja keeleõpetusmetoodika arengule. Sel eesmärgil on koostatud õppijakeele (ingl *learner corpus*) ehk vahekeele korpusi (*interlanguage corpus*), keeleõppeks sobivate tekstide korpusi (*corpora for learners*). Tänaseks on korpusepõhine vahekeele- ja veaanalüüs seotud uute õppematerjalide loomise (õpikud, grammatikad, sõnastikud) ja arvutipõhise keeleõppega (*computer assisted language learning*), samuti tõlkija- ja õpetajakoolitusega. Et kiirendada ja arendada õppurite keelematerjali kogumise, kirjeldamise ja uurimise kõiki etappe, on viimastel aastakümnetel loodud ja kasutusele võetud mitmesuguseid universaalsemaid ja kitsama suunitlusega korpusetarkvara lahendusi. Nii minnakse näiteks õppijakeele korpuste käsitsimärgendamiselt üle

* EVKK loomine on seotud järgmiste programmide ja sihtfinantseeritavate teemadega: sihtfinantseeritav projekt 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"; ETF-i grant nr 6151 "Koodivahetuse, eesti vahekeele ning lastekeele andmekorpuste koostamine ja üldkirjeldus" (2005–2008).
Autorid tänavad Lars Borinit Göteborgi ülikoolist heade soovituste ja informatsiooni eest.

poolautomaatsele. Automatiseeritud metoodikal põhinevad uurimused võimaldavad tänu muutunud materjalimahtudele saada usaldusväärsemaid tulemusi ja leida uut laadi informatsiooni vahekeele haruldasemate nähtuste kohta.

Käesolev artikkel tutvustab erinevat tüüpi õppijakeele korpusi, mis maailmas on loodud, ning Eesti vahekeele korpust (EVKK) nende taustal. Pikemalt peatutakse EVKK veaklassifikatsioonil ning kirjutatakse sellest, missugused teaduslikud ja rakenduslikud perspektiivid avanevad eesti keele kui teise keele uurimisel seoses korpuse kasutamisega.

Õppijakeel ehk vahekeel

Vahekeel kujuneb seoste alusel, mida õppija loob emakeele (K1) ja õpitava võõrkeele (K2) vahel (Selinker 1969, 1972, 1992, Corder 1981 jt), mõnikord ka juba omandatud võõrkeel(t)e, emakeele ning õpitava keele vahel (Michiels 1999, Cenos jt 2001, Hufeisen, Neuner 2003 jt). Vahekeelest on samuti räägitud kui sihtkeele variandist, mille kõrvalekaldumist normist mõjutavad vähemalt viis psühholingvistilist protsessi: 1) ülekanne K1 või K2, K3 jne põhjal; 2) ülekanne, mille õppija teeb õpetamise alusel; 3) õpistrateegiatest tulenev vigane keelekasutus; 4) suhtlusstrateegiatest lähtuv kõrvalekalle normist; 5) üldistamine ning reegli kasutamine selleks sobimatus kontekstis. Seega on vahekeel dünaamiline, tal puudub stabiilsus, tema areng on järk-järguline lähenemine sihtkeele normipärasele kasutamisele. Võib ka kujuneda olukord, kus vahekeel edasi ei arene. Ilmnevad fossiliseerumise ehk vea kinnistumise ning tagasimineku nähud: teatud suhtlussituatsioonides pöördub õppija vahekeele eelnevate etappide juurde ning hakkab tegema vigu, mida ta varem ei teinud (Tönshoff 1995: 5–6).

Õppijakeele korpustest maailmas

Õppijakeele korpused on maailmas uuemaid, kuid kiiresti levivaid korpuse tüüpe. Vanemaid neist on 1980. aastatel loodud Euroopa Teadusfondi teise keele andmepank (European Science Foundation Second Language Data Bank)[1]. Laiemalt hakati seda tüüpi korpusi rajama 1990. aastate algul, mitte-inglise sihtkeelega korpusi veelgi hiljem. Järgnevalt esitame olulisemad jooned, mis iseloomustavad erineva ülesehituse ja funktsiooniga õppijakeele korpusi.

Õppijakeele korpusi on loodud mitmesugustel eesmärkidel teoreetilisest uurimistööst kuni keeleõppe rakendusliku pooleni: näiteks eri tasemel keeleõppijate vahekeele tekstide alusel saab uurida teatud grammatilise struktuuri omandamist, õpikute ja sõnastike vastavust keeleõppijate tegelikele vajadustele (vt ka Kitsnik 2006: 94–99), õppijakeele korpust saab rakendada, koostades korpusepõhiseid lihtsamaid ja keerulisemaid keeleõppesüsteeme. Osa korpuseid on ette nähtud ennekõike vaid korpusepõhiseks uurimis- ja õppetööks, teised on planeeritud olema aluseks elektroonilistele keeleõppesüsteemidele. Arvutipõhise keeleõppe toetuseks on loodud mitmeid õppijakeele korpusi kasutavaid rakendusi: vigade identifitseerimine, viitamine grammatikareeglile või sõnastikule ja õpikeskkonnad. Õppijakeele korpuste koostamise põhimõtted ja analüüsivõimalused on mitmekesised.

[1] http://www.mpi.nl/ISLE/overview/Overview_ESFSLD.html (17.01.2007)

Muudeks olulisteks õppijakeele korpuste eristajateks on veel emakeelekõnelejate kontrollkorpuse olemasolu, allikmaterjali keel, kirjalike tekstide või suulise kõne kogumine, register ja tekstiliik. Mugava meetodina on viimasel ajal populaarsust võitnud elektrooniliste keeletestide korpuseks koondamine (vt Cobb 2003). Teiseks levinud tekstitüübiks on õppurite esseed. Samas leitakse, et õppijakeele alusuuringute tarvis on oluline kasutada mitte ainult kirjalikku, vaid ka suulist allikmaterjali, kuna võrreldes kirjalike tekstidega on suuline kõne vabam metalingvistilisest mõjust ning peegeldab õppija vahekeele arengut ja selle taga peituvaid mentaalseid protsesse paremini (vt Myles 2005: 375). Kirjaliku keele korpused toetavad eelkõige kirjutamis- ja lugemisoskuse uurimist ning arendamist. Leidub märgendamata ja märgendatud õppijakeele korpusi, viimased erinevad omakorda märgendamiseks valitud info ning märgendamise automaatsuse astme poolest.

Õppijakeele korpuste märgendamisel on praegusajal kasutusel ilmselt sama palju märgendamissüsteeme, kui palju on võõrkeele omandamise uurijaid. Sellele vaatamata võib täheldada selget liikumist standardiseerimise suunas. Parim näide on siinkohal CHILDES-i korpustevõrgustik[2], mida on arendatud 1980. aastate algusest saadik. Korpustevõrgustiku juurde kuuluva tarkvarapaketi märgendusmoodul on põhjalik, võimalusterohke ning tugeva tehnilise toega. Seda kasutatakse lapsekeele (emakeele) uurijate seas standardina, mida pidevalt uuendatakse ja täpsustatakse. Seetõttu kasutab tänaseks enamik suulise kõne õppijakeele korpusi CHILDES-i korpusevõrgustikku, mille märgendus on mugandatud CHILDES-i TEI standardile.[3] Korpuste loomiseks ja töötlemiseks on analoogselt CHILDES-iga väga võimalusterohke ka Esseni ülikoolis Raymond Hickey (2003) poolt koostatud 27 programmist koosnev üldine tarkvarapakett. WordSmithi korpusetarkvara leiab näiteks konkordantse, sagedasemaid sõnu tekstis (ingl *keywords*) jm.[4]

Maailmas on loodud nii avalikke huve kui kommertseesmärke teenivaid õppijakeele korpusi. Sellest sõltub ka korpuse ligipääsetavus, lisaks oleneb juurdepääsuvõimalus ka korpuse tehnilise teostuse valikust. Oluliseks eristavaks jooneks on ka erinevate korpuste mastaapsus ("suured" inglise sihtkeelel ja "väikesed" mitteinglise sihtkeelel põhinevad õppijakeele korpused).

Maailma suurimad õppijakeele korpused on enamasti inglise sihtkeelega: Cambridge Learner Corpus[5] (20 miljonit sõnet), Longman Learners' Corpus[6] (10 miljonit sõnet) ning International Corpus of Learners' English[7]. Viimane sisaldab 2 miljonit sõnet ja on samuti esimesi kirjaliku õppijakeele korpusi maailmas, selle on loonud 1990. aastate esimesel poolel Louvaini korpuslingvistika keskus eesotsas tuntud korpuslingvistide Sylviane Grangeri ja Fanny Meunieriga (vt Granger 2002). Väikesemahuliste korpuste (40 000–600 000 sõnet) hulka kuuluvad nt Antwerp Corpus of Institutional Discourse[8], prantsuse sihtkeelega FRIDA[9] ning Corpus of English by Japanese Learners[10]. Järgnevalt tutvustame neid korpusi lähemalt.

Cambridge Learner Corpus (CLC) ja **Longman Learners' Corpus** (LLC) – Cambridge´i ja Longmani õppijakeelekorpuste aluseks on kommertseesmärkidel loodud kirjastuste andmekogud, mille põhjal arendatakse keeleõppe

[2] CHILDES koosneb mahukast vestlustekstide andmebaasist, märgenduspõhimõtete süsteemist CHAT ning tarkvarapaketist CLAN, mis on mõeldud korpuse tekstide transkribeerimiseks, märgendamiseks ja analüüsiks. CHILDES-is on konkordantsileidjad, sagedusloendurid, sõnaliigimärgendajad paljudele keeltele, mida on üsna lihtne kohandada oma uuringuvajadustega. CHILDES-i aina laiemat levikut soodustab vaba ligipääs. http://childes.psy.cmu.edu/ (17.08.2006).

[3] TEI (Text Encoding Initiative) on rahvusvaheline juhendite kogu, mis võimaldab keeleuurijatel jt vormistada ning annoteerida erinevaid tekste. http://tei-c.org/ (17.08.2006).

[4] http://www.lexically.net/downloads/version4/html/index.html (27.08.2006)

[5] http://www.cambridge.org/elt/corpus/learner_corpus.htm (17.08.2006)

[6] http://www.longman.com/dictionaries/corpus/learners.html (17.08.2006)

[7] http://www.cecl.fltr.ucl.ac.be/Cecl-Projects/lcle/icle.htm (17.08.2006)

[8] http://www.contragram.ugent.be/newsle12.html#Acid (17.08.2006)

[9] http://www.latl.unige.ch/freetext/en/description.html (17.08.2006)

[10] http://leo.meikai.ac.jp/~tono/paper/crg.pdf (07.09.2006)

materjale. LLC ja CLC sisaldavad erineva emakeele ja keeletasemega õppijate vahekeeli. Vaba juurdepääs korpustele kahjuks puudub.

Saja erineva emakeelega õppurite kirjalikku inglise keelt peegeldav CLC koosneb tuntud standardsetest *online*-keeletestidest, mis jaotuvad allkorpusteks vastavalt õppija keeletasemele ja testi liigile. Märgendatavaks infoks on valitud grammatikavead. Korpuse töötlemise muudab mugavaks võimalusterohke statistikamoodul. Statistikat saab piiritleda õppija keeletaseme või testitüüpide järgi.

CLC-i andmeid kasutatakse rakendusuuringutes Cambridge'i ülikooli kirjastuse tarvis, et anda välja õppijate vajadustele paremini vastavaid õpikuid, keeletestideks valmistumise käsiraamatuid, sõnastikke ja grammatikaid.

International Corpus of Learners' English (ICLE) – rahvusvaheline inglise õppijakeelte korpus koosneb 19 erineva emakeelega üliõpilaste 500–1000-sõnelistest kirjalikest esseedest kogumahuga 2 miljonit sõnet. Allkorpuste mahuks on 200 000 sõnet. Sinna kuuluvad näiteks soome, poola, saksa, hiina, prantsuse ja leedu emakeelega kõnelejate korpused, lisaks inglise keelt emakeelena kõnelejate tekstidest koostatud testkorpus. Õppijate keelekasutust võimaldavad väga täpselt uurida erineva taustaga keelejuhtide ankeedid.

Korpuses on märgendatud ortograafiavead, leksikaalsed, leksikaalgrammatilised, grammatilised ja fraseoloogia vead. ICLE korpusele on loodud automaatne sõnaliigimärgendaja-lemmatiseerija.

Selle korpuse põhjal on viidud läbi teadusuuringuid paljudes riikides ja väga mitmekesistel teemadel (õppijakeele süntaks, sõnavara ja sõnasagedused, diskursuse ning stiili omapära, erinevate emakeeltega õppijate vahekeelte võrdlus).

Antwerp Corpus of Institutional Discourse (ACID) – Antverpeni institutsionaalse diskursuse korpus loodi 1990. aastate keskpaigas. ACID peegeldab edasijõudnute inglise keele kasutamist ametialases suhtluses, keskendudes võõrkeelse ning emakeelse, kirjaliku ning suulise diskursuse pragmaatiliste aspektide uurimisele ja võrdlemisele. Korpuses on 900 kirjalikku teksti, mis pärinevad äri- ja mittetulundusühingute ning akadeemiliste asutuste ametialasest suhtlusest. Korpus lubab võrrelda õppijakeelt emakeelekõnelejate omaga, samuti hollandlaste ja inglaste emakeelset kirjalikku ärisuhtluse keelt (vt Pelsmaekers 1997). Vaba ligipääs korpusele puudub.

Nii suulise kui kirjaliku diskursuse märgendamiseks on kohandatud CHILDES-i CHAT/CLAN-märgendussüsteem, mis võimaldab ka teha suurte tekstilõikude kvantitatiivset analüüsi. Korpuses on märgendatud kolme liiki andmeid: situatsiooniinfo, viisakusstrateegiate kasutamine ning semantiline-süntaktiline info (Pelsmaekers 1997). Situatsiooniinfo all märgendatakse alluvussuhted, tähistatakse ametikirja eesmärk konkreetses suhtlusprotsessis (algatus, küsimus jne), määratakse kommunikatsiooni domeen (nt äripakkumine) ja nn geograafiline info (nt asutusesisene *vs*. asutustevaheline kiri). Viisakusstrateegiad on märgendite moodulis olulisimad, rõhuasetus on ebamugavast situatsioonist väljatulemise strateegiatel. Süntaktiliste ja semantiliste märgendite moodul on vähem põhjalik ning samuti seotud viisakusstrateegiatega: kõneviis, tegumood, subjekti tüüp, põhiverbi tüüp, kõrvallause tüüp, modaalsed abisõnad jms.

FRIDA ja **FreeText**. FRIDA on prantsuse õppijakeelekorpus, FreeText on selle alusel loodud arvutipõhine keeleõppesüsteem. FRIDA koosneb 450 000 sõnest, mille moodustavad erinevate emakeeltega prantsuse keele õppijate kirjalikud

tekstid. Tekste püüti koguda võimalikult autentsetest keelekasutussituatsioonidest, kus õppijad käituvad loominguliselt (seminarid, interaktiivsed õppetegevused, tekstimõistmisülesanded ning kontrollitud ja vaba tekstiloome ülesanded). FRIDA korpuses on vead märgendatud ning iga vea tähtsust on hinnatud. Vaba ligipääs korpusele ja keeleõppesüsteemile puudub.

FreeText valmis 2003. aastal ülikoolide ja ärisektori koostöös Euroopa Liidu rahvusvahelise projektina. Osalesid 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. FreeTexti veadiagnoosimise süsteemi on treenitud FRIDA märgendatud osa peal, mille maht on 300 000 sõnet. FreeTextis töötavad kõrvuti automaatse veaotsinguga ka teised loomuliku keele automaattöötlusvahendid: süntaksianalüsaator, lausestruktuuri äratundja, lause reformuleerija ja tõlkija, kõnesüntesaator. Veadiagnoosimissüsteemis on eraldi välja toodud ortograafia-, grammatika- ja semantikavead. FreeText on esirinnas arvutipõhise keeleõppe arendamisel: siia kuuluvad interaktiivne õpikeskkond, mis võimaldab õppida prantsuse keeles koostama viit tüüpi kirjalikke tekste ning mida õpetaja saab vastavalt vajadusele kohandada; grammatikakontrollija-speller; grammatika käsiraamat, mille konkreetsetele lõikudele tehakse õpikeskkonnas viiteid; sõnastikud.

FreeTexti automaatse veadiagnoosimise süsteemi aluseks on mõjustus- ja sidusteoorial põhinevad tehnikad (vt Vandeventer Faltin 2003). Analüsaator vaatleb õppijakeele kõiki tasandeid, parandab testülesannete vead, annab tagasisidet ning viiteid konkreetsetele grammatikareeglitele e-käsiraamatus. Grammatikakäsiraamat on koostatud veastatistika alusel ning see täieneb pidevalt õppuritele antava tagasiside põhjal. FreeTexti loomine on andnud olulise panuse keeleõppeteooria ning õpetuspraktika arengusse, avardunud on arusaamad vahekeele korpuste koostamise ja arendamise võimalustest, veadiagnostikast ning keele automaattöötlusvahendite kasutamisest vigade leidmisel ja märgendamisel.

Corpus of English by Japanese Leaners (CEJL) – jaapani-inglise õppijakeele korpuse maht on ligi miljon sõnet ja sisaldab jaapanlaste kirjaliku ning suulise keelekasutuse näiteid. CEJL koosneb neljast allkorpusest: 1) ingliskeelsed kirjalikud tekstid (12–19-aastaste noorte jutustav ja argumenteeriv proosa); 2) ingliskeelne suuline materjal (argumenteerimine ja koomiksi kirjeldus); 3) jaapanikeelsed kirjalikud tekstid (jutustav ja argumenteeriv proosa); 4) inglise keele kui võõrkeele õpikutekstide korpus (kontrollkorpus). Vaba juurdepääs korpusele puudub.

CEJL-i alusel on uuritud vahekeele arenemise mustreid ning võrdlevalt ka emakeele ja õppijakeele korpuseid. Korpuse töötlemisel on kasutatud erinevat tarkvara, näiteks WordSmith ja MonoConc Pro. Märgendamisel on kolm aspekti: sõnaliigid, vead ja teksti metainfo. Veamärgendusel eristatakse nii lingvistilist klassifikatsiooni (nt vealiigitus keeletasandite järgi – fonoloogiast teksti ja diskursuse vigadeni; sõnaliikide vead; eri tasandi moodustajate vead; vealiigitus grammatikakategooriate alusel – aeg, tegumood, loendatavus, transitiivsus, referendi ja keeleüksuse vahelise suhte iseloom, mille hulka on arvatud ka kollokatsioon) kui vealiigitust sihtkeele modifitseerimisviisi alusel (nt ärajätt ja lisamine). Ärajätt on siin erinevalt ellipsist ja nullelementidest mittegrammatiline nähtus: **He'll pass his exam and I'll ø too* 'Ta läbib eksami ja ma ø ka', lisamine on keelereegli järgimine vales situatsioonis.

Veel eristab CEJL-i märgendus sihtkeele modifitseerimisviisi järgi ebakorrektset vormimoodustust regulariseerimise, irregulariseerimise või topeltmarkeerimise tõttu, õigete vormide vale kasutust ja järjestamist. Segunemisvigadeks peetakse niisuguseid nähtusi nagu ühtesulamine, kontaminatsioon, ristassotsiatsioon ja hübridiseerimine: **according to Erica's opinion* '*Erica arvates järgi' – õiged on: *according to Erica* / *in Erica's opinion* 'Erica arvates / Erica arvamuse järgi'.

Teksti metainfo sisaldab teavet ülesannete tüüpidest (loominguline tekst, suuline kõne, paaristöö, jutustamine), õppeasutusest, kust tekst on pärit, läbitud keeleõppe mahust, tööle antud hindest.

Mitte-inglise sihtkeelega korpustest

Inglise õppijakeele korpusi on loodud tunduvalt enam kui muudel sihtkeeltel põhinevaid, kuid oma võimaluste ja mahu poolest on mitte-inglise sihtkeelega korpused (ingl *non-English learner corpora*) sageli huvipakkuvamad. Pikaajalised kogemused mitte-inglise õppijakeele korpuste arendamisel on näiteks Skandinaavias, kus juba aastatel 1973–1980 Björn Hammarberg jt lõid ligi 112 000-sõnelise rootsi sihtkeelega **SSM korpuse**[11] (Svenska som Målspråk), mis koosneb eri emakeelte ja keeletasemetega õppijate rootsikeelsetest esseedest.

Stockholmi ülikooli lingvistikateaduskonnas on välja töötatud **Andraspråkets StrukturUtveckling korpus** (ASU)[12] – õppijatelt erinevatel õppeetappidel kogutud sihtkeele struktuuri arengu korpus ehk nn longituudkorpus. ASU sisaldab kirjalikku ja suulist materjali rootsi kui kolmanda keele õppijatelt, kajastades nende keelekasutuse ajalist arengut. Korpuse kogumaht on 490 000 sõnet, sisaldab ka kontrollkorpust. ASU on loodud erilaadsete võrdlevate uuringute tarvis ning õppijakeele veaanalüsaatorite arendamiseks.

Rootsis on erinevad õppijakeele korpused koondatud **SVenska AndraspråksTExter** ehk **SVANTE korpusesse**[13] (Göteborgi ülikool), mis kuulub üldisesse rootsi keelepanka (Språkbanken). SVANTE korpuse tekstid on lemmatiseeritud ning sõnaliikide osas märgendatud, korpust täiendatakse pidevalt.

Alates 2001. aastast kasutatakse SVANTE korpust Stockholmi kuningliku tehnoloogiainstituudi ja Stockholmi ülikooli koostööprojektis **CrossCheck**[14] (Svensk grammatikkontroll för andraspråksskribenter, rootsi õppijakeele grammatikakontrollija), mille raames tehakse ulatuslikku arendustööd rootsi keele kui teise keele elektrooniliste õppevahendite koostamisel. CrossChecki olulisemaid tulemusi on veadiagnoosimise süsteem (vt Bigert 2005), mis arvestab sõna kasutuskonteksti iseärasustega.

Rootsis on koostatud ka prantsuse, saksa ja hispaania sihtkeelega õppijakeele korpusi. Soomes pole seni mitte-inglise õppijakeele elektroonilistele korpustele rõhku pandud, eelistatud on tõlke- ja paralleelkorpuste arendamist, mille uurimisest saadud tulemusi on kasutatud nii tõlkijakoolituses kui keeleõppe eesmärkidel. Huvi õppijakeele korpuste loomise vastu on ilmnenud alles viimasel ajal: III Virsu konverentsil Joensuus tutvustas Jarmo Jantunen (2007) laiahaardelist rahvusvahelist projekti soome õppijakeele korpuse loomisest Oulu ülikoolis.

Norra õppijakeele korpus **Language learner corpus of Norwegian as a second language**[15] koosneb immigrantide keeletestide arhiivist. Korpuses on nii

[11] Infot korpuse kohta vt http://www.ling.su.se/DaLi/research/xcheckpres.html (01.02.07).
[12] http://www.ling.su.se/staff/ham/projects.html (07.01.2007)
[13] http://www.svenska.gu.se/~svelb/svante/ (07.01.2007)
[14] http://www.csc.kth.se/tcs/projects/xcheck/index-en.html (07.01.2007)
[15] http://decentius.aksis.uib.no/corpus/askdemo-home.xml (01.02.2007)

sõnaliigi-, süntaktiline kui vealiigimärgendus. Veaanalüsaator lisab veamärgendi, annab hinnangu vea raskusastme kohta ja teeb veaparanduse. Teistest Euroopas kõneldavate keelte õppijakeele korpustest võib esile tuua rahvusvahelise itaalia õppijakeele korpuse VALICO[16] ning FLLOC-andmebaasi[17] (koondab kuut prantsuse sihtkeelega suulist õppijakeele korpust).

Eesti vahekeele korpuse üldtutvustus

Tallinna Ülikooli üld- ja rakenduslingvistika õppetooli Eesti vahekeele korpus[18] kuulub väikeste mitte-inglise õppijakeele korpuste alla, sisaldab vene emakeelega eesti keele õppijate kirjalikke töid, on koostatud ja elektrooniliselt töödeldud kindlal uurimistöö eesmärgil. Korpuse olulisimaks töövahendiks on konkordantsiprogramm, mis annab erinevate kriteeriumide alusel paindlikud otsinguvõimalused. Et luua konkordantsiprogramm, tuli eelnevalt defineerida vea mõiste, keelevigade liigitamise alus ning sellest lähtuvalt jõuda mitmemõõtelise lingvistilise veaklassifikatsioonini (vt EVKK mitmemõõtelisest lingvistilisest veaklassifikatsioonist).

EVKK on kõigile veebis kättesaadav ükskeelne avatud korpus, millesse võib tekste jätkuvalt lisada. 2006. aasta lõpus on EVKK osaliselt märgendatud tekstikogu, mis sisaldab osaliselt tasakaalustatud alg-, kesk- ja kõrgtaseme tekste erineva sotsiaalse taustaga keeleoskajatelt kogumahus 459 731 sõnet (märgendatud 75 749 sõnet, millest 8123 olid vigase keelekasutuse näited). Korpuse alusel saab uurida nii korrektset kui vigast õppijakeelt. Korpust saab kasutada tekstiarhiivina (ilma märgenduseta) või märgendatud variandina, kus keeleviga on välja toodud ja lingvistiliselt tõlgendatud. Tekstikogust saab leida samaliigilisi vigu, näha neid kitsamas kontekstis (süntagma), ± 1–2 lauses või konkreetses terviktekstis. Jälgida saab ka mingi sõna õigeid ja vigaseid esinemisnäiteid ning sama vealiiki ühes tekstis, ühes tekstitüübis või kõikides korpuse tekstides. Samuti on võimalik keelevigade infot seostada õppuri sotsiaalse päritolu, staatuse, hariduse, soo, vanuse ja emakeelega ning keelevaldamise tasemega (sarnaselt Euroopa Nõukogu keeleoskustasemete süsteemiga A-, B- ja C-tase).

Järgnevalt anname ülevaate EVKK teksti- ja informandivalikust, samuti korpuse väljatöötamise etappidest.

EVKK metainfo kohta annab jooksvalt teavet korpuse statistikalehekülg. Näiteks autori keeletaseme seisukohalt sisaldab korpus hetkel kõige enam kesktaseme tekste (52,4%), järgnevad alg- (22,5%) ja kõrgtaseme (25,1%) tekstid. Autorite sooline kuuluvus (64,7% naised, 35,3% mehed) ja haridustase (keskharidus 67,4%, põhiharidus 18,7% ja kõrgharidus 13,4%) ei ole tasakaalus. Disproportsioon on nähtav samuti tekstide autorite vanuses (52,9% informantidest on kuni 18-aastased) ja sotsiaalses taustas (49,5% üliõpilasi, 27% õpilasi, 16% teenistujaid jne). Autorite valikul on teadlikult piirdutud kahe Eestimaa regiooniga, kus on kõige rohkem venekeelset elanikkonda: 59,9% korpuse tekstide autoritest on pärit Tallinnast või Harjumaalt ja 29,1% Ida-Virumaalt. Samas tuleb märkida, et informantide hulka, kes on end kirja pannud Tallinna/Harjumaa elanikuna, võib olla sattunud ka Ida-Virumaalt ja teistest Eesti piirkondadest pärit inimesi (Tallinnas õppimas või täiendkoolitusel).

Teksti tüüpidest sisaldab EVKK kõige rohkem abivahendeid kasutamata kirjutatud esseid (46%), vastuseid küsimusele (17,1%) ja kirju (8,6%). Eelistatud

[16] www.bmanuel.org/projects/br-HOME.html (29.12.2006)
[17] http://www.flloc.soton.ac.uk (4.01.2007)
[18] http://evkk.tlu.ee (26.02.2007)

on loomingulist laadi töid. Et aga õppetöös tuleb ette ka muid kirjaliku kontrolli vorme (referaat, tõlge, grammatikaharjutus, muu – 10,7%), siis ei olnud põhjust neid õppijakeele tekste korpusest välja jätta.

Metainfo statistika põhjal saab teha mõningaid järeldusi korpuse tasakaalustatuse kohta. Samas aga ei tasu tasakaalustatuse eesmärki korpuse tegemisel absolutiseerida. Võõrkeeli omandatakse enamasti noores eas, rahule jäädakse põhiliselt alg- ja kesktasemega, naised õpivad meestest meelsamini jne.

EVKK arendamine on toimunud etapiti. Esimene etapp oli tarkvara valik seatud eesmärkidest lähtuvalt. Sobivat vabavara EVKK töötlemiseks mitmemõõtelise lingvistilise veaklassifikatsiooni alusel (vt Eslon 2006b: 15–17, 19–20) ei olnud võimalik leida. Seetõttu tuli luua originaalne tehniline teostus (insener Vahur Rebas), mille aluseks võeti Tallinna Ülikooli haridustehnoloogia keskuse tarkvara ja avatuse põhimõte.

Järgmise etapina asuti korpust käsitsi märgendama. Töö käigus selgus, et nii tehnoloogilist poolt kui esialgset lingvistilist veaklassifikatsiooni tuleb järk-järgult täpsustada. Selles tööfaasis ei saa kasutada olemasolevaid eesti keele tehnoloogilisi ressursse (vt ülevaade Muischnek jt 2003). Näiteks morfoanalüsaator analüüsib grammatiliselt õigeid vorme, kuid vigade leidmise ja märgendamise eesmärgil on see kasutu.

Korpuse loojad on otsustanud avada 500 000 sõnest koosneva märgendatud korpuse, milles sisalduvat materjali saab kasutada nii teadusliku uurimistöö eesmärgil kui õpetajakoolituses ja praktilises keeleõppes. Kasutajal on võimalik luua korpusse oma töökeskkonda ning arendada veamärgendust ja tekstivalikut vastavalt kitsale uurimisteemale.

Hetkel pole korpuse kodulehel avalikkusele veel kättesaadavad päring, statistika, veaklassifikatsioon ning märgendamata ning märgendatud tekstid, küll aga saab EVKK ja õppijakeele korpustega seotud infot. Korpuse kõiki mooduleid on kasutatud Tallinna Ülikooli eesti filoloogia osakonna üliõpilaste ning magistrantide õppetöös ja uurimistegevuses.

EVKK mitmemõõtelisest lingvistilisest veaklassifikatsioonist

EVKK-s on keeleviga mõistetud grammatikareeglile või kommunikatiivsele eesmärgile mittevastava keelekasutusena, mille hulka ei kuulu väsimusest ja hooletusest põhjustatud eksimused ning keelevääratused. Samuti ei vaadelda vigadena keelelisest kreatiivsusest põhjustatud variatsioone, mille eesmärk on erinevate keelesiseste võimaluste mänguline kasutamine mõtte väljendusrikkamaks edastamiseks. Veaklassifikatsioonist on välja jäetud ka keele kultuurilised ja psühholoogilised aspektid.

Veaklassifikatsiooni loomisel on arvestatud keelendi õigsust edastatava tähenduse seisukohast (semantika), keelendi vormilist õigsust (grammatika) ning kasutuse sobivust (pragmaatika). Tegu on lingvistilise veaklassifikatsiooniga, milles lähtutakse keelest kui paradigmaatiliste ja süntagmaatiliste seoste süsteemist (vt tabel 1). Paradigmaatilised seosed on hierarhilised, iga alamtaseme keelend sisaldub igas kõrgemas tasandis ning iga kõrgema tasandi keelend hõlmab alamatasemelisi (grafeem, morfeem, sõna, sõnaühend, lause, tekst) ja vastupidi. Süntagmaatilised seosed on lineaarsed, neis kombineerub leksika-, morfoloogia- (leksikaalgramma-

tika, morfonoloogia, morfoloogia, morfosüntaks) ja lausetasandi (süntaks, registri valik ja pragmaatika) reeglistik. Minimaalne märgendusüksus on sõne, maksimaalne lause. Lausest suuremate üksuste märgendamine (nt märgendid "teksti sidususvahendite kasutamine" ning "teksti sõnastuse ja registrivaliku vastavus autori pragmaatilisele eesmärgile") on taandatud miinimumini.

Keelesüsteemi süntagmaatilise ja paradigmaatilise telje ristumisel tekib kahemõõteline lingvistiline veaklassifikatsioon (vt Hufheisen 1991, Michiels 1999), mille põhjal saab eristada 18 veaklassi (vt tabel 1).

Tabel 1. Keelevigade lingvistiline klassifikatsioon

Süntagmaatika / Paradigmaatika	**Semantika**	**Grammatika**	**Pragmaatika**
Tekst	1	2	3
Lause	4	5	6
Sõnaühend	7	8	9
Sõna	10	11	12
Morfeem	13	14	15
Grafeem	16	17	18

Iga veaklass sisaldab erineva hulga konkreetseid liike ja alamliike, mis on omakorda vastastikku paradigmaatilistes ja süntagmaatilistes seostes. Vealiike ja alamliike on EVKK veaklassifikatsioonis kokku 170. Ülemliikide hulka kuuluvad leksikaalsed ja sõnatuletusvead; leksikaalgrammatilised, morfonoloogilised, morfoloogilised, morfosüntaktilised, süntaktilised ja kommunikatiivsed vead. Iga ülemliigi vead jagunevad alamliikideks. Näiteks leksikaalgrammatiliste vigade alamliikidena on välja toodud tegevuse piiritletus/piiritlematus, transitiivsus/intransitiivsus, tegevuslaadi väljendavad verbisufiksid, analüütilised verbid, afiksaaladverbide kasutamine, noomeni tuletusliidete kasutamine. Iga alamliik võib omakorda jaguneda: näiteks tegevuse piiritletuse/piiritlematuse all on välja toodud omastavaline/nimetavaline ja osastavaline sihitis; ainult osasihitist võimaldavad verbid; ainult täissihitist võimaldavad verbid; konteksti vahendid, mis määravad tegevuse piiritletuse või piiritlematuse; kvantiteedisõnade kasutamine tegevuse piiritlemise vahendina jne. Üldisele veaklassifikatsioonile on lisatud ka märgend "Proovi kätt", millega märgendatakse neid vigu, mida erinevatel põhjustel pole olnud võimalik selgelt määratleda.

Veaklasside, -liikide ja alamliikide väljatoomine ning alamliikide jagunemine veelgi kitsamateks vea ilminguteks annab võimaluse kirjeldada viga mitmemõõteliselt, rääkida veast mitte ainult kui tagajärjest, vaid välja tuua ka vea tekkepõhjusi (vt Eslon 2006b). Need kaks poolt – põhjus ja tagajärg – on vigade interpreteerimisel omavahel lahutamatult seotud. Ühe või teise märkimata jätmine tekitab ebaadekvaatseid tõlgendusi, mis omakorda võivad üle kanduda õppeprotsessi, veateraapiasse ja -diagnostikasse, luues nähtumusel põhinevaid formaalseid või suisa valesid seoseid. Seetõttu toetub EVKK märgendussüsteem veaklassi, -liigi ja alamliikide vahelistele seostele. Vea tekkepõhjuste väljatoomisel arvestatakse ka

K1 ja K2 vaheliste sümmeetria, asümmeetria ja analoogiaseoste olemasoluga (vt Eslon 2006a: 22). Näiteks:

(1) kui sõna *viskuma* asemel on ekslikult kasutatud sõna *viskama*, siis võib viga tõlgendada mitme veaklassi omavaheliste seoste alusel: veaklass 10 (semantika kokkupuutepunkt sõna tasandiga), veaklass 11 (grammatika kokkupuutepunkt sõna tasandiga), veaklass 13 (semantika kokkupuutepunkt morfeemi tasandiga), veaklass 14 (grammatika kokkupuutepunkt morfeemi tasandiga), veaklass 16 (semantika kokkupuutepunkt grafeemi tasandiga) ja veaklass 17 (grammatika kokkupuutepunkt grafeemi tasandiga). Teisisõnu: semantika- ja grammatikaviga on eristatav grafeemi tasandil (16 ja 17), kuid põhjustatud eksimusest morfeemi tasandil, kuna ilmselt ei ole õppija omandanud *u*-refleksiivi semantikat (13) ja grammatikat (14), mille tagajärjeks on kahe erineva sõna semantiline (10) ja grammatiline (11) äravahetamine lauses.

(2) Ainsuse osastava käände ekslik lõpp *-k* sõnas *arvutik* (*Ma õpen vell arvutik*) võib olla interpreteeritud kui grammatikaviga, mis avaldub grafeemi tasandil (veaklass 17 – grammatika kokkupuutepunkt grafeemi tasandiga, grammatikaviga grafeemi tasandil) ning on põhjustatud oskamatusest moodustada ainsuse osastavat käänet (veaklass 14 – grammatika kokkupuutepunkt morfeemi tasandiga, grammatikaviga morfeemi tasandil). Viga on võimalik mõtestada ka häälduspärase kirjaviisina, mille võib olla tinginud õppija metakeeleline seisukoht – "kirjutan nagu kuulen". Selgelt on see nähtav ka näiteks sõna *prägu* (õige: *praegu*) kasutamisel (*Ja prägu ma kirjutan sulle kirja eesti keelest tunni*). Tegelikult kajastub hääldus pärases kirjaviisis vene ja eesti häälikusüsteemide erinevus, mis põhjustab nihkeid häälikute tajumisel ning ilmneb häälduses ja õigekirjas.

(3) Lauses *Oman kõrgharidust ja neljaaastane töökogemus sissetuleva turismi alal* on allakriipsutatud sõnaühend saanud korpuses mitu märgendit. Kõigepealt on esile toodud sõnaühendi grammatikaviga (veaklass 8 – grammatika ja sõnaühendi kokkupuutepunkt): ainsuse nimetava käände asemel peab kasutama ainsuse osastavat (*Oman .. nelja-aastast töökogemust* pro *neljaaastane töökogemus*). Seejärel on märgendaja näinud lause tasandi grammatikaviga (veaklass 5 – grammatika ja lause kokkupuutepunkt): liidetud on normaallause ja omajalause, milles sisalduvatel predikaatverbidel on erinev argumendistruktuur (*kes omab mida?*, kuid *kellel on mis?*). Õppijakeeles on aga toimunud üldistamine, mille tulemusel nimetatud kaks struktuuri on ühendatud normaallause malli alusel. Järelikult saab välja tuua süntaksivea alamliigi – lauseliikmete ärajätmine (*Oman kõrgharidust ja mul on nelja-aastane töökogemus*). Teise interpretatsiooni kohaselt võib siin näha kahe objektnoomeni ühildumisviga arvus ja käändes (*Oman kõrgharidust ja nelja-aastast töökogemust*). Analüüsitav viga on märgendatud ka verbirektsiooni veana (veaklass 8), objektnoomeni käände veana (veaklass 14).

Toodud näited kinnitavad ammuteada seisukohta vea interpretatiivsest iseloomust (Corder 1981) – tavapäraselt saab üks keeleviga ikka mitu märgendit ning see ei johtu veaklassifikatsiooni puudulikkusest, vaid on vea olemuse väljendus. Kui loomuliku keele süsteemi lingvistilise kirjeldamise puhul räägitakse interpretatiivsetest

kategooriatest ja keelendite keerulisest funktsionaalsest potentsiaalist (vt Бондарко 1996), siis sama kehtib ka vahekeele süsteemi ja selles ilmnevate vigade suhtes – interpretatiivsus on veale iseloomulik tunnus. Järelikult peab vea olemusega adekvaatne klassifikatsioon olema mitmemõõteline: iga sügavama tasandi viga on iga kõrgema tasandi konkretiseering vähemalt kolmes aspektis (semantika, grammatika, pragmaatika). Hetkel on EVKK vealiikide määramise aluseks ligi 300 tunnust. Niisugust veaklassifikatsiooni võib visualiseerida vealiikide vastastikuste sõltuvuste puuna, tänu millele võime näha vigade hierarhiat ning kirjeldada viga ja selle tekkepõhjusi tunduvalt avaramalt, samas ka täpsemalt kui seni on tehtud (nt Stemberger 1985: 33–39, Lähdemäki 1995, Pool, Vaimann 2005).

Lisaks võimalusele õppijakeelt süsteemselt uurida saab mitmemõõtelise veaklassifikatsiooni alusel parema ülevaate ka sellisest tähtsast mõjurist sihtkeele omandamisel nagu lähte- ja sihtkeele sisesed keelendite ja kategooriate vahelised üleminekualad ehk leksikaalgrammatiline perifeeria, milles ilmnevad keelendite sekundaarsed funktsioonid ning millele tugineb keeleline kreatiivsus. Just kategooriate üleminekualadel ja keelenditevahelistes seostes realiseerub nende funktsionaalne potentsiaal, võimalus olla kasutatud samades või erinevates sekundaarsetes funktsioonides ning samades või erinevates kontekstitüüpides. Seetõttu ongi oluline uurida ühelt poolt K1 ja K2 grammatilist perifeeriat ning teisalt otsida kahe erineva keelesüsteemi vahelisi seoseid, mis on aluseks emakeele positiivsele mõjule teise keele omandamisel (vt Kaivapalu 2005: 23–38).

Samas on ka vahekeele analüüs näidanud, et "vead tekivad eelkõige sel juhul, kui õpitava ja emakeele keelendeid kasutatakse sekundaarsetes funktsioonides" (vt Гак 2001: 18). Järelikult aitab seoste ning üleminekualade leidmine erinevate veaklasside, liikide, alamliikide jne vahel meil paremini mõista nii õppijakeele olemust kui vigade tekkepõhjusi. Seoste ilmnemisel peaks olema võimalik teatud vealiikide järgi prognoosida teiste nendega enamseotud vigade esinemise tõenäosust.

Järgnevalt mõned näited sellest, millist osa etendavad lähte- ja sihtkeele vahelised seosed, samuti mõlema keele süsteemi perifeersed nähtused vene emakeelega eesti keele õppijate vigade tõlgendamisel. Näited (4–8) on leitud EVKK märgendatud korpuseosast päringumooduli abil.

(4) Otsingule "põhitähenduse viga" andis päring hetkel näiteid 26 tekstist, milles see vealiik on märgendatud. Näidete hulgast valisime sõna *värske* ebakorrektse kasutuse (*On vaja, et meie riietus on <u>värske</u> ja puhas*), mis on seotud selle sõna sekundaarse tähendusega, eelkõige õppija emakeeles. Vea täpsustusena on lisatud märgend "sõnade semantiline ühildamatus". Põhjus: vene keeles on sõna *свежий* 'värske' sekundaarseid tähendusi 'äsja pestud ja triigitud, puhas' (rõivastest kõnelemise kontekstis: *на нём свежая рубашка / рубашка не первой свежести = tal on seljas puhas / mitte eriti puhas särk*). Eesti keeles see sekundaarne tähendus puudub – *riietus* ei saa olla *värske*, kuna need kaks sõna on omavahel semantiliselt sobimatud. Küll aga võib riietusese olla *värskelt triigitud*, s.t *äsja*. Näide kuulub veaklassi 7 (semantika kokkupuutepunkt sõnaühendi tasandiga), vealiik – leksikaalne viga, alamliik – väljendustava vastu eksimine, alamliigi jaotused – sõnade semantiline ühildamatus ja kollokatsioon. Toodud näite alusel saab väita, et K1 ja K2 üleminekualaks ning samas neid keeli siduvaks nähtuseks on

adjektiivide *värske* ja *свежий* semantiline ühildamatus/ühildatus substantiividega *riietus* ja *одежда*.

(5) Otsingule "paronüümi kasutamine", mis kuulub leksikaalsete vigade alamliiki "põhitähenduse viga", leidis päring vastuseks 30 teksti. Analüüsime tekstikatket *Ilmuvad igasuguseid kirjanduserühmid näiteks: "Tarapita" – mis oli pooleli poliitiline, ja pooleli kirjanduslik ühend* (õige: *ühendus*), *veel üks kirjanduse rühm "Siuri" (Maria Under, Tuglas, Alle ja teised.*). Õppijakeeles on segi läinud sihtkeele häälduselt lähedased, kuid vormilt ja tähenduselt erinevad sõnad *ühend* ja *ühendus*, mille tulemusena tekib paronüümia (veaklass 10 – semantika kokkupuutepunkt sõna tasandiga). Samas kannab see juhtum ka märgendit "afiksaalse sõnatuletuse viga (nimisõnaliite *-us* kasutamine)" – veaklass 11 (grammatika kokkupuutepunkt sõna tasandiga) ja veaklass 14 (grammatika kokkupuutepunkt morfeemi tasandiga). Vea tekkepõhjus peitub sihtkeele sõnade *ühend* ja *ühendus* tähendus- ja vormierinevuste mittenägemises, mille tulemusel õppija ei adu piiranguid nende semantilises ühildatuses/ühildamatuses (vrd *keemiline ühend = химическое соединение*, kuid *kirjanduslik ühendus = литературное объединение*). Positiivne ülekanne oleks võimalik seoste loomisel lähtekeelega, kus *ühend* ja *ühendus* vasteks on nii vormilt kui põhitähenduse poolest omavahel erinevad sõnad *соединение* ja *объединение*.

(6) Otsingule "samasse mõistepesasse kuuluva sõna kasutamine" leidis päring vastuseks 8 teksti, mille hulgast valisime järgmise näite: *Minu arvates see oleks tore reis. Võib* (õige: *saab*) *tutvuda uute inimestega ja veeta aega rõõmsalt ja kasulikult*. Lauses esile toodud viga kuulub klassi 10: liik – leksikaalne viga, alamliik – põhitähenduse viga, mille jaotuseks on "viga samasse mõistepesasse kuuluva sõna kasutamisel". Täpsustusena on lisatud märgend "viga modaalsõnade kasutamisel". Teise interpretatsiooni kohaselt võib antud juhtumit siduda veaga sõna tähendusvarjundite eristamisel (klass 10, liik – leksikaalne viga, alamliik – tähendusvarjundi viga). Põhjus peitub ühelt poolt kahe sõna semantika osalises kattumises, teiselt poolt erinevustes kontekstuaalsetes piirangutes. Modaalverbi *võima* tähendusväljas saab eristada vähemalt nelja võrdväärset komponenti: 1. '(hüpoteetiliselt) võimalik olema', 2. 'reaalne olema, reaalselt võimalik olema', 3. 'lubatud olema', 4. 'suuteline olema, oskama'. Verbi *saama* põhitähendus on aga mittemodaalne – 'omandama, kätte saama'. Modaalverbide klassiga on ta seotud sekundaarse tähenduskomponendi 'reaalne olema, reaalselt võimalik olema' vahendusel (vrd *Siin on väga kärarikas. Lähme ärklikorrusele, seal saame puhata*), millele lisandub ajatähenduse varjund – lähitulevikulisus. Kuna analüüsitava tekstikatke autor kirjutab reaalsest võimalusest tutvuda lähitulevikus uute inimestega ning veeta aega rõõmsalt ja kasulikult, siis lause mõtet arvestades tuleks sõna *võib* asemel kasutada *saab* – see, millest räägitakse, saab olema ja saab tulema. Tegu on veaga, mitte aga sünonüümsusest tuleneva variatiivsusega. Sellest annab järeldada, et modaalverbide *võib* ja *saab* kontekstiseotus tuleneb keelesisestest varjatud protsessidest, mille aluseks on keelendite sekundaarsed tähendused ja funktsioonid keelesüsteemi leksikaalgrammatilises

perifeerias. Vene emakeelega õppijale ei ole eesti keele perifeersete nähtuste alla kuuluv erisus sugugi kergesti arusaadav, kuna vene avaratähenduslikul modaalverbil *мочь* 'võima' on väga rikas semantiline potentsiaal, mis ei pane verbi kasutamisele kontekstuaalseid piiranguid. Sama laieneb ka modaalsõnale *можно* 'võib'.

Lisame näiteid teistelt keeletasanditelt.

(7) Morfosüntaksi all on ühe alamliigina välja toodud viga "*ma*-infinitiivi kasutamine seoses faasiverbiga": *Seal me hakkame elada ja ööbime* (õige: *hakkame elama ja ööbima*). Teise interpretatsiooni kohaselt on siin morfoloogiaviga: alamliik – indikatiivi oleviku ajavormi kasutamine (*Seal me elame ja ööbime*). Analüüsitava näite puhul on oluline, kuidas teksti autor mõtestab sündmuse ajalist kulgu. Sõlmküsimuseks saab tulevikulisuse semantika väljendamine. Eesti keele analüütilise verbikonstruktsiooniga *hakkame elama* ja vene keeles sellele vastava imperfektiivse aspekti analüütilise tuleviku vorm *будем жить* edastavad meile arusaama kaugemas tulevikus toimuma hakkavast tegevusest. Sel puhul on eesti tegevusverb alati *ma*-infinitiivi vormis ja vene tegevusverb imperfektiivses aspektis. Sisuliselt on siin tegu keeltevahelise üks-ühese vastavusega. Sama nähtus on jälgitav ka eesti keele sünteetilise verbi oleviku vormi *vaatame* ja vene keele perfektiivse aspekti oleviku-tuleviku vormi *посмотрим* vahel, sest mõlemad väljendavad lähitulevikulisust, tegevuse vahetut järgnemist kõnemomendile, nt *vaatame, mis teha annab* = *посмотрим, что можно сделать*; *astun kohe tema juurest läbi* = *сразу же зайду к нему* jne. Olenevalt lause semantilisest, samuti ajadeiktiliste sõnade ning muude rõhumarkerite olemasolust kontekstis võib ajatähendusele lisanduda ka modaalne tähendusvarjund – kõnemomendile järgnev tegevus on reaalselt võimalik (*Nad lähevad ju omavahel tülli* = *Они же поссорятся*). Kuna eesti ajavormidel ja vene aja-aspektivormidel on tulevikulisuse semantika edastamisel ühesugune funktsionaalne potentsiaal, siis kõneleb see üks-ühesest seosest nende vahel. Seetõttu on vene emakeelega õppijal kaugema ja lähitulevikulisuse väljendamisel võimalus toetuda keeltevahelisele korrelatiivsusele, mis samas aitab realiseerida K1 keelekompetentsis sisalduvaid teadmisi, luues tingimused positiivse ülekande tekkimiseks.

(8) Otsing "moodustajate süntaktilised funktsioonid" tõi päringus välja 32 teksti; üks selle vealiigi alamliikidest on "sihitise viga", nt *Me ka kasutame pesuvahendid* (õige: *pesuvahendeid*). Eesti keeles sõltub objektnoomeni käändevalik mitmest tegurist: näiteks objekti määratlematusest (umbmäärane kogus), kirjeldatava situatsiooni imperfektiivsusest/perfektiivsusest.

Objekti määratlematus/määratletus ning selle olulisus aspektisituatsiooni mõtestamisel on vene lähtekeelega õppija jaoks erandlik nähtus, kuna tema emakeeles sõltub sihitise kasutamine transitiivverbi rektsioonistruktuurist, on seotud käändekategooriaga ning alles seejärel räägitakse objekti olemusest (elus/eluta, haaratud tegevusest osaliselt/täielikult, otsene/kaudne suunaobjekt). Tegu on erinevustega keelemeeles, mis kajastub ka grammatikakirjeldustes ja terminoloogias. Katsed

ühitada eesti ja vene keele objektiuuringutes otse- ja kaudsihitist on seetõttu esile kutsunud vastuväiteid: eesti keeles puudub kaudobjekti süntaktiline omapära (vt Erelt 2002). Seega on eesti keele kui teise keele õppes mõttekas mitte arutada selle üle, kas objekt on määramata/määratud, vaid toetuda objektinoomeni käändevaliku õpetamisel mõlema keele jaoks universaalsetele momentidele nagu kirjeldatava situatsiooni imperfektiivsus/perfektiivsus.

Kuna näitelause (8) semantika viitab tegevuse pidevusele ja seega imperfektiivsele iseloomule, siis kasutatakse objektnoomenit reeglina osastavas käändes. Sama viga on märgendatud ka leksikaalgrammatilisena, alamliik – tegevuse piiritletuse/piiritlematuse väljendamine. Selgituseks: kui subjekt midagi pidevalt kasutab, siis on tegu piiritlemata tegevusega, mida võib mõtestada ka subjektile omase harjumusena (meil on harjumus või komme kasutada pesuvahendeid). Kui aga millegi kasutamine on subjekti jaoks ammendunud, siis kirjeldatakse tegevust piiritletuna. Selle väljendamiseks sobivad eesti keeles ühendverbid, antud juhul *ära kasutama*. Siinkohal võib näha selgepiirilist korrelatsiooni K1 ja K2 vahel: eesti keeles on tegevuse piiritlematuse/piiritletuse ja imperfektiivsuse/perfektiivsuse edastamisvahend objektnoomeni käändevorm (partitiiv vastandatuna nominatiivile/genitiivile); vene keeles aga vastandus imperfektiivne/perfektiivne aspekt. Peale selle on eesti keele objektnoomeni käände kasutamine seotud ka sünteetilise/analüütilise verbivormi valikuga. Neist esimene kirjeldab piiritlemata tegevust – *kasutan, kasutasin pesupulbrit*; *olen, olin kasutanud pesupulbrit*, teine aga tegevuse piiritletust – *kasutan, kasutasin pesupulbri ära*; *olen, olin pesupulbri ära kasutanud*. Võrdluses vene keelega ilmneb vastavus eesti keele sünteetilise verbi ja vene keele imperfektiivse aspekti ning eesti analüütilise verbi ja vene perfektiivse aspekti vahel. Järelikult on lause aspektuaalse semantika interpreteerimine K2-s tihedalt seotud K1 keeleteadmistega ja universaalset laadi keeltevaheliste korrelatsioonidega.

Kokkuvõtvalt: mitmemõõtelise veaklassifikatsiooni alusel saab uurida sihtkeele reeglite vale kasutamist ja ekslikke ülekandeid õppijakeeles süsteemselt, tuues välja teise keele omandamiseks olulised universaalsed nähtused, mis tingivad vajaduse uutel alustel rajanevate ainekavade, õpikute, õppematerjalide loomiseks. Selles osas on mõttekas võrrelda ka õppijakeelt ja sõnasagedust loomuliku keelekasutuse ja sõnasagedusega.

EVKK ja pedagoogiline grammatika

Tavapäraselt on pedagoogilise grammatika loomisel tuginetud kontrastiivgrammatikale, mille ideoloogia on seotud teesiga K1 interferentsi otsustavast mõjust K2 omandamisel. Nii on rõhutatult esile toodud K1 negatiivne mõju K2 omandamisele. Selle teooria seavad kahtluse alla paljud vahekeele uuringud, samuti eelnenud veaanalüüsi näited (vt EVKK mitmemõõtelisest lingvistilisest veaklassifikatsioonist).

Vahekeelt võib vaadelda seoste võrgustikuna lähte- ja sihtkeele vahel, mida õppija ei loo kontakteeruvate keelte kontrastsuse põhjal, vaid vajadusest K2-st aru saada, otsides seejuures kokkulangevusi, lähedasi ja analoogseid nähtusi. Ning kui midagi jääb seletamata, seosed eelneva ja järgneva vahel on loomata, siis tekitab õppija neid ise – lihtsatest assotsiatsioonidel põhinevatest ülekannetest abstraktsete

analoogiateni (nt Ehala 2000: 28). Seetõttu peaks pedagoogilise grammatika loomisel tingimata arvestama vahekeele veaanalüüsi tulemusi: missugused vead ja kui sageli ette tulevad (vea esinemissagedus), kuidas vead on omavahel seotud (seoste statistika), kuidas õppijakeele vead on seotud sihtkeele autentse sõna- ning vormikasutuse ning -sagedusega, sõnaühendite moodustamise ja muu sarnasega. Alles seejärel tasub asuda otsima K2 omandamise tegelikke raskuspunkte. Seejuures on muidugi oluline, et vahekeele uurimine oleks laiapõhjaline ning tugineks erinevate lähtekeeltega õppijate sihtkeele kasutusnäidetele nii kirjalikus kui suulises kõnes. Niisugust seisukohta on varemalt väljendanud nt F. Myles, kes leiab, et õppijakeelt saab analüüsida kahel viisil: 1) seades kitsamaid eesmärke ja võrreldes õppijate vahekeelt sihtkeele korrektse grammatikaga ning 2) seades avaramaid eesmärke ja uurides õppija vahekeelt kui iseseisvat terviklikku lingvistilist süsteemi, milleks aga on vaja võrrelda erinevaid emakeeli rääkivate õppijate keelekasutust liikumisel sama sihtkeele poole (Myles 2005: 378, vt ka Granger 1998).

Võib minna ka teist teed ja võrrelda vahekeele veaanalüüsi tulemusi K1 ja K2 vaheliste seoste grammatikaga ehk korrelatsioonigrammatikaga (mille kohta vt Eslon 2006a: 24, Eslon 2006b: 17–21). Vastupidiselt kontrastiivgrammatikale ei tugine korrelatsioonigrammatika keelekontrastidele, vaid püüab näha tüpoloogiliselt erinevate keelte vahel üks-üheseid vastavusi ning samalaadsust, mille tulemusena rakendub lähtekeele analoogiapõhine positiivse ülekande mehhanism. Just seetõttu ei saa keeleõppest välja jätta K1 ja K2 süsteemide kõrvutavat kirjeldust, kuid erinevalt kontrastiivgrammatikast tuleb see viia teistlaadi teoreetilis-metodoloogilisele alusele – põhjuslike seoste tuvastamisele lähte- ja sihtkeele vahel (vt Eslon 2006a: 17, 19–20). Pedagoogilise grammatika koostamise eesmärgil on mõttekas vahekeele veaanalüüsi tulemusi võrrelda K1 ja K2 korrelatsioonigrammatikaga ning leida vahekeelt ja korrelatsioonigrammatikat ühendavaid momente. Tegelikkuses tähendab see keeleteooria, rakenduslingvistika ja ka ainedidaktika integreerimist keeleõppe eesmärgil (vt ka Hudson 2004).

Rääkides seni koostamata eesti keele kui võõrkeele pedagoogilisest grammatikast, on käesoleva artikli autoritel eelnevalt olnud mõte seostada selle loomine EVKK-põhise vahekeele analüüsi tulemuste ja korrelatsioonigrammatikaga. EVKK veaanalüüsiga on seotud empiiriline suund pedagoogilise grammatika loomisel, korrelatsioonigrammatika väljatöötamisega aga teoreetiline suund. Muidugi oleks ideaalne luua eesti-*x*-keele korrelatsioonigrammatika, mis teatud mõttes oleks universaalse grammatika näide. Esialgu jääb see ilmselgelt üle jõu käivaks ülesandeks, kuna eeldab ühe võimaliku väljundina keelelise käitumise modelleerimist erinevates keeltes ning suhtlussituatsioonides. Esialgne töö piirneb vene emakeelega õppijate vahekeeletekstides sisalduvate vealiikide, alamliikide jne määramise ning nendevaheliste seoste väljaselgitamisega, mis sagedusstatistika rakendamisel aitab jõuda vigade tekkepõhjusteni. Paralleelselt toimub uurimistöö eesti-vene korrelatsioonigrammatika koostamiseks. Aluseks on võetud funktsionaalsemantilised väljad (vt Eslon 2006a: 20–21).

Pedagoogilisel grammatikal on võõrkeeleõppes keskne koht: grammatika teeb valiku keeleainesest, mis ainekavas järjestatakse keelendite sageduse, produktiivsuse ja omandamise raskuse alusel. Pedagoogilise grammatika põhjal on võimalik üles ehitada süsteemne keeleõpe, koostada selle läbiviimiseks vajalik ainekava, kirjutada õpikud, luua sõnastikud ja muud õppeprotsessi toetavad vahendid.

Vahekeele korpuse alusel tehtud veastatistika ning erinevatesse veaklassidesse ja -liikidesse kuuluvate vigade omavaheliste seoste võrgustiku kindlakstegemine aitab välja töötada eesti keele õppeks vajaliku reeglite süsteemi.

Kirjandus

Bigert, Johnny 2005. Automatic and Unsupervised Methods in Natural Language Processing. Doctoral Thesis. Stockholm. http://www.nada.kth.se/~johnny/docs/thesis.pdf (9.01.2007).

Бондарко Александр Владимирович 1996. Проблемы грамматической семантики и русской аспектологии. Санкт-Петербург: Изд-во Санкт-Петербургского ун-та.

Cenoz, Jasone; Hufeisen, Britta; Jessner, Ulrike (Eds.) 2001. Looking Beyond Second Language Acquisition. Studies in Tri- and Multilingualism 6. Tertiärsprachen und Mehrsprachigkeit Bd. 6. Tübingen: Stauffenberg Verlag.

Cobb, Thomas 2003. Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. – Canadian Modern Language Review 59 (3), 393–423.

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

Ehala, Martin 2000. Second language learners´ impact on the structure of Estonian. – Kiira Allikmets (Ed.). Languages at Universities Today and Tomorrow. Proceedings of the Methodology Conference of the Language Centre, 19-20 May 2000. Tartu, 20–32.

Erelt, Mati 2002. Hierarhiatest tüpoloogias. – Renate Pajusalu, Ilona Tragel, Tiit Hennoste, Haldur Õim (toim.).Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu: TÜ Kirjastus, 34–40.

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

Eslon, Pille 2006b. Eesti vahekeele korpusest korrelatsioonigrammatikani. – Helle Metslang, Margit Langemets (toim.), Maria-Maren Sepper (keeletoim.). Eesti Rakenduslingvistika Ühingu aastaraamat 2. Estonian Papers in Applied Linguistics 2. Eesti Rakenduslingvistika Ühing. Tallinn: Eesti Keele Sihtasutus, 11–24.

Гак Владимир Григорьевич 2001. Семасиологический функциональный подход и типология функций. – Теоретические проблемы функциональной грамматики. Материалы Всероссийской научной конференции, Санкт-Петербург, 26–28 сентября 2001 г. Отв. ред. Александр Владимирович Бондарко. Санкт-Петербург: Наука, 17–19.

Granger, Sylviane 1998. The computerized learner corpus: A versatile new source of data for SLA research. – Sylviane Granger (Ed.). Learner English on Computer. London, New York: Longman, 3–18.

Granger, Sylviane 2002. A Bird's-eye view of learner corpus research. – Sylviane Granger, Joseph Hung, Stephanie Petch-Tyson (Eds.). Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching. Amsterdam/Philadelphia: John Benjamins, 3–33.

Hickey, Raymond 2003. Corpus Presenter. Software for Language Analysis (With a Manual and A Corpus of Irish English as Sample Data) + CD-ROM. Amsterdam, Philadelphia: John Benjamins. http://www.uni-essen.de/~lan300/HICKEY.htm (27.08.2006).

Hudson, Richard 2004. Why education needs linguistics (and vice versa). – Journal of Linguistics 40 (11), 105–130.

Hufeisen, Britta 1991. Englisch als erste und Deutsch als zweite Fremdsprache. Empirische Untersuchung zur fremdsprachlichen Interaction. Frankfurt/M etc: P. Lang.

Hufeisen, Britta; Neuner, Gerhart (Eds.) 2003. Mehrsprachinkeitkonzept – Tertiärsprachen – Deutsch nach Englisch. European Centre for Modern Languages. Strasbourg: Council of Europe Publishing. http://www.ecml.at/documents/pub112G2003.pdf (9.03.2007).

Jantunen, Jarmo 2007. Oppijansuomen piirteitä korpusvetoisesti. – III kansainvälinen Virsu-konferenssi 19.-20. tammikuuta 2007 Joensuussa. http://www.joensuu.fi/suomi/virsu/virsabstraktit007.html (6.02.2007).

Kaivapalu, Annekatrin 2005. Lähdekieli kielenoppimisen apuna. Maisa Martin, Pekka Olsbo, Marja-Leena Tynkkynen (toim.). Jyväskylä Studies in Humanities 44. Jyväskylä: Jyväskylän yliopisto.

Lähdemäki, Eeva 1995. Mikä meni pieleen? Ruotsinkielisten virheet suomen ainekirjoituksessa. Fennistca 11. Åbo: Åbo Akademis tryckeri.

Michiels, B. 1999. Die Rolle der Niederländischkenntnisse bei Französischsprachigen Lernern von Deutsch als L3: Eine empirische Untersuchung. – Zeitschrift für Interkulturellen Fremdsprachenunterricht 3 (3). http://www.spz.tu-darmstadt.de/projekt_ejournal/jg-03-3/beitrag/mich1.htm (6.02.2007).

Muischnek, Kadri; Orav, Heili; Kaalep, Heiki-Jaan; Õim, Haldur 2003. Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Urve Talvik (toim.). Tallinn: Eesti Keele Sihtasutus.

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

Pelsmaekers, Katja 1997. Dutch, English and English L2 business language in contrast: Working with the ACID corpus. – Contragram (Quaterly newsletter of the Contrastive Grammar Research Group of the University of Gent) 12. http://www.contragram.ugent.be/newsle12.html#Acid (6.02.2007).

Pool, Raili; Vaimann, Elle 2005. Vead kõrgtasemel eesti keele kõnelejate kirjalikus keelekasutuses. – Margit Langemets (koost.), Maria-Maren Sepper (toim.). Eesti Rakenduslingvistika Ühingu aastaraamat 1. Estonian Papers in Applied Linguistics 1. Eesti Rakenduslingvistika Ühing, Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus, 115–137.

Selinker, Larry 1969. Language transfer. – General Linguistics 9 (2), 67–92.

Selinker, Larry 1972. Interlanguage. – International Review of Applied Linguistics 10, 209–231.

Selinker, Larry 1992. Rediscovering Interlanguage. New York: Longman.

Stemberger, Joseph P. 1985. The Lexicon in a Model of Language Production. New York, London: Garland Publishing Inc.

Tönshoff, Wolfgang 1995. Ausgewählte Forschungsergebnisse und Denkanstösse für die Unterrichtspraxis. – Fremdsprache Deutsch (Zeitschrift für Praxis des Deutschunterrichts). Sondernummer 1995: Fremdsprachlerntheorie, 4–15. http://www.edition-deutsch.de/fremdsprache/pdf/fd_s095.pdf (05.10.2005).

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

Pille Eslon (Tallinna Ülikool) on uurinud vene keele funktsionaalset grammatikat, üld- ja kõrvutava keeleteaduse probleeme, modaalsust, aspektuaalsust, infinitiivi.
pille.eslon@tlu.ee

Helena Metslang töötab Integratsiooni Sihtasutuses, teaduslikeks huvialadeks on süntaks, eriti lauseliikmetevahelised üleminekualad, ning õppijakeele korpused.
helena.metslang@meis.ee

LEARNER LANGUAGE AND ESTONIAN INTERLANGUAGE CORPUS

Pille Eslon, Helena Metslang
Tallinn University / Non-Estonian's Integration Foundation

The current article introduces the design, aims and future applications of the Estonian Interlanguage Corpus (EIC) in the background of other learner language corpora in Europe and Asia. Estonian Interlanguage Corpus is an annotated monitor corpus of learner language texts, being created at the chair of General and Applied Linguistics at Tallinn University. It comprises written Estonian language texts, mainly by Russian speakers. The objective of EIC is to support the development of pedagogy and research and of acquisition of Estonian as a second language.

EIC is a free online corpus which will by the end of its first stage of development have 500 000 strings of error-annotated texts plus a larger volume of non-annotated texts. The most important tool of EIC is the concordancer which gives flexible search options by various criteria. The specially created error classification has 170 tags and is multidimensional (there are the syntagmatic and paradigmatic axes). To the highest level of tags there belong lexical errors, derivational errors, lexicogrammatical, morphophonological, morphological errors, morphosyntactic, syntactic and communication errors and unspecified errors (ambiguous, doubtful errors).

The results of EIC error analysis are a good basis for writing a systematic pedagogical grammar for Estonian as a second language. Analyzing errors statistically and determining the network of links between the errors will help to describe better the system of rules necessary for learning Estonian as a second language. After creating the pedagogical grammar it is possible to build on it syllabi, courses, textbooks, dictionaries and other means to support the process of learning Estonian as a second language.

Keywords: corpus linguistics, interlanguage, learner language corpus, pedagogical grammar