TALLINNA ÜLIKOOLI
EESTI KEELE JA KULTUURI INSTITUUDI
TOIMETISED 11

KORPUSUURINGUTE METODOLOOGIA JA MÄRGENDAMISE PROBLEEMID

Toimetanud Pille Eslon ja Katre Õim

Tallinn 2009

Tallinna Ülikooli Eesti Keele ja Kultuuri Instituudi toimetised 11
Publication of the Institute of Estonian Language and Culture 11
Tallinn University

Postiaadress / Contact information:
Eesti Keele ja Kultuuri Instituut / Institute of Estonian Language and Culture
Tallinna Ülikool / Tallinn University
Narva mnt 25
10120 Tallinn
ESTONIA

ekki.toimetised@tlu.ee
www.tlu.ee

ISSN 1736-4221
ISBN 978-9985-58-668-6

SISUKORD

EKSPERIMENTAALSE METOODIKA KASUTAMISEST EESTI KEELE OMANDAMISE ALASTES UURINGUTES

Reili Argus

Ülevaade

Eesti keele omandamise alastes uuringutes ei ole eksperimentaalset metoodikat seni kuigivõrd kasutatud. Seetõttu antakse siinses kirjutises esmalt lühike ülevaade teiste keelte omandamise uurimuste aluseks olnud peamistest katsetüüpdest ning kommenteeritakse neid isiklikest kogemustest lähtuvalt. Seejärel kirjeldatakse kaht keeleomandamiskatset, võrreldakse väikesearvulise vastanute valimi põhjal saadud tulemusi ning osutatakse katsete koostamise ja läbiviimisega seotud probleemidele, samuti olulisematele üksikasjadele, millega katse koostaja ning uurija arvestama peavad.

Võtmesõnad: eesti keele omandamine, lastekeel, psühholingvistilise katse koostamine, katse ülesehitus[1]

[1] Artikli valmimisel on toeks olnud ETF-i grant "3–7aastaste laste keelelise arengu leksikaalsed ja grammatilised seaduspärad".

1. Keeleomandamisuurimustes enam kasutatud katsetüüpidest

Psühholingvistilisi katseid rakendatakse keeleomandamise uurimustes enamasti kahel juhul: esiteks siis, kui keelenähtus on tavakasutuses harv ning selle omandamist lihtsalt ei saa pikiuuringuga kogutud materjali põhjal jälgida, ja teiseks siis, kui on oluline täpsemalt teada saada, mis tingimused või missuguste tingimuste muutumine mõjutab ühe või teise keelendi omandamist ja kuidas see toimub. Sageli ongi mingi keeleomandamiskatse eesmärk selgitada välja, missugune keelestruktuuri osa on mingis vanuses lapse jaoks raske. Näiteks on teada, et alakõnega laste keelesüsteemis on teatud kindlad piirkonnad, kus tehakse palju vigu, ja mingid kindlad grammatilised kategooriad, mille omandamine on raske (vt nt Leonard 2003: 791). Pikiuuringuga kogutud materjaliga võrreldes on katse eeliseks ka see, et ühe ja sama keelenähtuse kohta käivat keelematerjali on võimalik saada nii eri vanuses kui ka suuremalt arvult samaealistelt lastelt.

Keeleomandamisuuringutes kasutatud katsetüübi valik on sõltunud eelkõige uurimuse suunast ja uurija teoreetilisest taustast (nt sellest, kas uurija on psühholoog, neurolingvist, sotsiolingvist või klassikaline lingvist). Keeleomandamiskatsete rakendamist võib pidada levinumaks neuro- ja psühholingvistikas, teistest enam kasutavad katseid generatiivse suuna esindajad. Viimastel aastatel on katsed rakendust leidnud ka kognitiivset suunda pooldavate lingvistide töödes.

Üks esimestest keeleomandamisuurimuste katsetüüpidest on **reaktsioonikiiruse mõõtmine** (ingl k *Reaction Time Studies*). Reaktsioonikiiruse mõõtmist on bioloogia-alastes uuringutes kasutatud juba alates 19. sajandi keskpaigast (Kosinsky & Cummings 1999), keeleomandamises leidis see katsetüüp rakendust

mõnevõrra hiljem (Grane & Thornton 1998: 66). Reaktsioonikiiruse mõõtmisega soovitakse leida vastuseid küsimusele, kas erinevate keelestruktuuride mõistmiseks kulub ühepalju aega. Katse puuduseks on peetud seda, et reaktsioonikiirust mõjutavaid tegureid on liiga palju, alates katsealuse soost, paremavasakukäelisusest kuni hingamisrütmini välja (Kosinsky & Cummings 1999: 80-82).

Teise tüübina on keeleomandamisuurimustes kasutatud katseid, kus laps peab mingis lauses väljendatud **tegevust jäljendama** (ingl k *The Act-Out Task*) (Grane & Thornton 1998: 67). Sellistel katsetel on hulk puudusi, millest olulisim on vast asjaolu, et kõikide keelestruktuuride mõistmist ei ole võimalik mitteverbaalselt (jäljendamisega) demonstreerida – ka siis, kui laps mõistab keelendi tähendust, ei pruugi ta suuta seda mitteverbaalseid vahendeid kasutades demonstreerida.

Üheks visuaalse stiimuli põhiseks katsetüübiks on ka **pildivalikukatse** (ingl k *Picture Pointing Task*). Selle käigus kasutatakse mingile keelendile vastavat pilti koos nn segajatega ehk piltidega, mis konkreetsele testitavale keelendile ei vasta, ning laps peab piltide hulgast üles leidma õige (vt nt Vinnitskaya & Wexler 2001).

Eespool nimetatud kolme põhilist katsetüüpi on kasutatud keelestruktuuride mõistmise uurimiseks. Tunduvalt keerukam on aga tekitada olukord, kus laps peab mingit struktuuri ise kasutama, nt looma lause või vormi. Üks lihtsamaid, kuid samas võimalustelt kõige piiratumaid vahendeid on eelkõige alakõne kindlakstegemisel kasutatav **imiteerimisülesanne**, kui lapsele öeldakse järjest sõnu, vorme ja lauseid, mida tal tuleb järele korrata. Sellise katse rakendamisel on saadud näiteks tulemuseks, et alakõnega lapsed kordavad sõnavorme järele nii, et muutemorfoloogilised tunnused-lõpud jäävad sageli ära

ja ette öeldud liitlausest on korduses järele jäänud ainult lause esimese pool.

Senistes keeleomandamisuurimustes on kõige sobivamaks peetud kahte liiki katseid: mõistmis- ja loomistesti. **Mõistmistestis**[2] (ingl k *Truth Value Judgement Task*) tuleb hinnata mingi lause tõesust; **loomistestis** (ingl k *Elicited Production Task*) peab aga laps poolikut lauset jätkama (vt nt Grane & Thornton 1998). Algselt kasutati neid katsetüüpe generatiivse taustaga keeleomandamisuurimustes, nüüdseks on need levinud väga paljude eri teoreetilise taustaga lingvistide töödesse.

Nimetatud katsetüüpide kõige suuremaks eeliseks võib pidada seda, et mõistmis- ja loomistesti kombineerides saab paralleelselt uurida nii ühest ja samast struktuurist arusaamist kui ka selle struktuuri iseseisvat loomist. Mõistmistesti puhul kasutatakse sobiva situatsiooni tekitamiseks tavaliselt mingit visuaalset stiimulit, nt pilte, viimasel ajal aina sagedamini lühikesi videofilme. Demonstreeritava situatsiooni kohta öeldakse lapsele lause, mille õigsust ta peab hindama. Ka loomistesti puhul kasutatakse situatsiooni visualiseerimist ning öeldakse ette lause algus, mida laps peab jätkama. Seejuures on oluline, et situatsioon oleks võimalikult täpselt piiritletud, et lapsel tekiks soov oma lausejätkus kasutada just soovitud vormi või kategooriat.

Kõnealuste testide puhul on uurija jaoks positiivne see, et katse läbiviija kontrollib olukorda. Võrreldes pikiuuringuga saab sellise katse abil uurida mõne kindla konstruktsiooni, vormi või kategooria mõistmist ja selle kasutusest arusaamist, samuti saada ettekujutus neid protsesse mõjutavatest teguritest (nt kuivõrd mõjutab küsilausete omandamist see, kas

[2] Siinses kirjutises on kasutatud terminit *katse* ülemmõistena ning teminit *test* katse ühe poole kohta ehk alammõistena.

küsimus on subjekti või objekti kohta, kas subjekt on ainsuses või mitmuses, kas objekti käändevahelduse tähenduse tajumine sõltub ka objektnoomeni morfofonoloogilistest omadustest, sh vältevaheldusest jne). Lisaks on võimalik tekitada situatsioon, mida laps tajub eelkõige kui mängu, mitte kui hindamist. Puuduseks võib pidada suurt ajakulu (ühe lapse individuaalsele testimisele kulub vähemalt pool tundi), samuti asjaolu, et katse läbiviija küll kontrollib olukorda, kuid ei saa seda alati teha absoluutse järjekindlusega. Sagedasim probleem seisneb selles, et katse kavandamisel on raske luua konteksti, kus soovitud vorm või sõna oleks lause jätkamise ainuvõimalik variant. Näiteks võib tuua objekti käändevahelduse eksperimendi (vt Argus 2009), kus loomistesti lauset *Punases pluusis tüdruk …* (*... sõi küpsise ära*) kippusid lapsed jätkama *… sai kiiremini valmis*, kusjuures soovitud objektnoomenit, mille käände omandamist uuriti, lapse vastuses ei leidunudki. Selle konkreetse olukorra vältimiseks tuli lastele lause algusest ette öelda pikem osa, milles sisaldus ka verb: *Punases pluusis tüdruk sõi … .* Vaid sel juhul kasutati 98% juhtudest objektnoomenit. Kui aga sihtkeelendiks on mingi muu verbivorm, siis on vastavat vormi tingiva konteksti loomine oluliselt raskem. Kogemuste põhjal võib väita, et lastele detailsemate vastamisjuhiste andmine ei ole alati otstarbekas, näiteks on viie- ja kuueaastastel lastel vastamist suunavaid juhiseid üsna raske mõista. Juhis „Räägi nüüd mulle tüdrukust, kasuta sõnu *sööma* ja *kook*“ võib osutuda liiga keerukaks ning tajudes, et ülesanne on ülemäära raske, jäävad lapsed lihtsalt vait ega soovigi enam katses osaleda. Kui katse eesmärk on panna laps looma mingit kindlat vormi, mitte lekseemi, siis üks võimalustest on esitada sihtlekseeme saatetekstis võimalikult sageli. Näiteks, kui soovitakse panna laps kasutama objekti käändena genitiivi vormi sõnast *sild*, siis juhatatakse situatsioon sisse lausetega „Räägi mulle nüüd

tüdrukust. Tal tuleb teha *sild*". Lapsele jääb meelde, et see, millest tal rääkida tuleks, on just *sild*. Kindla vormi omandamise uurimisel on oluline, et lapsele suunatud saatetekstis sihtvormi ette ei anta ning seega ei tohiks situatsiooni kirjeldamiseks kasutada saatelauset „Tüdruk peab ehitama *silla*". Samas hoolikalt läbi mõeldes on saatelauseisse alati võimalik leida situatsiooni konkretiseerivaid mittesihtvorme, nt „Tal tuleb ehitada üks *sild* (ainsuse nominatiiv)".

2. Keeleomandamiskatsete ülesehitusest ja katsete koostamisega seonduvast

Enamasti kontrollitakse katse läbiviimise protseduuri ladusust ja ülesehituse eesmärgipärasust täiskasvanute kontrollgrupiga; kui katsega kavatsetakse määrata keelelist mahajäämust, siis moodustub teine kontrollgrupp normaalse keelelise arenguga lastest. Vähemalt 20 katseisikust[3] koosneva kontrollrühma kasutamise eesmärk on teha kindlaks, kas väljatöötatud katsega on üleüldse võimalik sihtkeelendi omandamist uurida ja testküsimustele 100% ootuspäraseid vastuseid saada. Alles siis, kui täiskasvanute vastuste põhjal võib öelda, et kõikidele küsimustele, testlausetele jne on reaalne saada ootuspäraseid vastuseid, võib katset hakata läbi viima ka lastega. Selleks sobiva kontrollrühma moodustamisel tuleks arvestada võimalikult paljusid tegureid, näiteks võiks kontrollrühmas olla võrdselt mehi ja naisi, samuti võiks varieeruda katsealuste vanus. Üsna sage on olukord, kus kontrollrühmana kasutatakse (mugavust silmas pidades) üliõpilasi, kuid keeleoman-

[3] Rahvusvahelise projekti COST A33 „Crosslinguistically Robust Stages of Language Development" raames välja töötatud katsete puhul leppisid 28 eri keele omandamise uurijad omavahel kokku, et kontrollgrupi piisav suurus on 20 inimest.

damisalastes eksperimentides tuleks hoiduda lingvistiliste erialade üliõpilastest. Kogemus on näidanud, et filoloogid suhtuvad katsesse ülemääraste ootustega. Näiteks võiks tuua keeleomandamiskatse (vt Argus 2009), kus kontrollrühmana kasutatud üliõpilaste kommentaaridest katse lõpus selgus, et nad ei uskunud sageli, nagu võiksid küsimused ja seega ka vastused olla nii lihtsad, kui need esmapilgul tundusid. Kahtlustati, et „seal peab kindlasti olema veel midagi, mingid nipiga küsimused". Võimalik, et just seetõttu oli nende vastuste hulgas rohkem mitteootuspäraseid vastuseid kui näiteks 6aastaste laste vastustes.

Kui aga kontrollrühma tulemused ei ole saja protsendi ulatuses ootuspärased, tuleb kaaluda kõigepealt seda, kas keelestruktuur, mida katsega uurida kavatsetakse, on täiskasvanute kõnes n-ö stabiilne, kas ei esine näiteks varieeruvust lokaalsete või sotsiaalsete murrete kasutajate keelekasutuses, kas mitte ei ole tegemist keelesüsteemi osaga, mis on keeles parasjagu muutumas (vt objekti käände kohta Ehala 2007; Argus 2009). Kui on selge, et nähtus ei ole täiskasvanute keelekasutuses ebastabiilne, siis tuleb keskenduda katse ülesehitusele ja analüüsida kontrollgrupi vastustes iga konkreetse katseüksuse tulemusi eraldi. Katse mõne testüksuse taustasituatsioonis või saatelauseis võib olla midagi sellist, mis tingib mitteootuspärase vastuse. Üks võimalus teha kindlaks, mis konkreetselt mõnes kindlas testüksuses mitteootuspäraseid vastuseid tingib, on koostada katse lisavariant, milles problemaatiliste küsimuste juures on ka vastusevariandi valikut selgitavad küsimused. Näiteks võiks olla objekti käändevaliku katse testüksus, kus on tegemist lõpetatud tegevusega ja katseisikule esitatakse lause *Punases pluusis tüdruk pani pusle kokku*, mille ta peaks õigeks tunnistama. Sellisele testüksusele tuleks vastuse juurde lisada selgitus, miks katseisik arvab, et vastus

ei ole õige, andes vastusevariantidena 1) Jah; 2) Ei. – Miks ei? Lisaküsimus *Miks ei?* selgitab välja eeskätt katse mittekeelelise materjali puudujäägid, sest vastajal on võimalus viidata, et näiteks pildil (või filmis) ei olnud hästi näha, kas tegelane lõpetas oma tegevuse või mitte, või ei olnud näiteks aru saada, et just see konkreetne tegelane kõnealust tegevust tegi vms.

2.1. Katseisikute vanus

Üldiselt on väidetud, et katsetes saab kasutada vähemalt kahe ja poole aasta vanuseid lapsi, sest nooremad ei suuda katse protseduuri järgida ega saa aru, mida neilt täpselt oodatakse (Crain & Thornton 1998: 145). Lisaks on kõnealune vanus keeleomandamiskatseteks sobiv ka seetõttu, et 2,5 aastat on olnud piiriks, millest alates on lingvistid lõpetanud spontaanse kõne regulaarse lindistamise ning vanemate laste kohta pikiuuringuga saadud keelematerjali lihtsalt ei leidugi. Keelelise mahajäämuse uurimisega seoses tuleb arvesse võtta, et alakõnega ja normaalse keelelise arenguga laste keeleomandamise võrdlemiseks ei panda katsealuseid tavaliselt mitte vanuserühmadesse (ei lähtuta laste bioloogilisest vanusest), vaid nad jaotatakse rühmadesse spontaanse kõne **väljendite keskmise pikkuse** (VKP, ingl k *MLU*) alusel. Nii saadakse näiteks grupid, kus laste VKP on suurem kui 1, suurem kui 2 jne.

Nelja- kuni seitsmeaastaste eesti laste testimisel kogutud tähelepanekute põhjal võib väita, et nelja-aastastel lastel on katse ülesehitusest ja neile esitatud ootustest mõnikord raske aru saada. Vastuseid, mis näitavad, et laps ei saa aru, mida ta täpselt tegema peab, on nelja-aastaste laste vastuste hulgas pea poole rohkem kui viieaasaste laste vastuste hulgas (vt Argus 2009). Kuni viieaastaste laste jaoks paistab eriti raske olema lausete jätkamine loomistestis. Seega võib väita, et mida nooremad on katseisikud, seda enam tähelepanu tuleb pöörata

treeningüksustele. Loomistesti puhul võib abiks olla n-ö kolmas tegelane, näiteks käpiknukk või mängukaru[4], kelle pooleli jäänud lauseid peab laps jätkama. Treeningsituatsiooni mängulisus ja vabadus on tihedalt seotud edasiste ülesannete mõistmisega.

2.2. Katseüksuste arv ning järjestamine

Katseisikute hulk ja katseüksuste üldarv on omavahel tihedalt seotud. Kui katses kontrollitakse ühe kindla keelestruktuuri omandamist, siis sõltub katseüksuste arv katsealuse keeleüksuse olemusest, eelkõige sellest, kui palju on tingimusi, mille puhul teatud struktuuri kasutamist või mõistmist kontrollitakse. Näiteks kui tahetakse teha kindlaks, millal omandavad eesti lapsed grammatilised ajad, täpsemalt (liht)mineviku, oleviku ja tuleviku, siis tuleb pidada silmas, et hästi tasakaalustatud katses peaksid kõik kolm aega esinema vähemalt kolm korda. Teisisõnu: kui katse on üles ehitatud nii, et laps peab hindama lausete tõesust, siis mõistmistestis peaks nii minevikku, olevikku kui ka tulevikku tähistava situatsiooni kohta esinema olevikuvorm vähemalt üks kord; samuti on minevikuvormiga jne. Järelikult on sellise katse minimaalne testüksuste arv üheksa ja kui uurija soovib midagi väita näiteks eesti laste minevikuvormide mõistmise kohta, siis peab ta silmas pidama seda, et sellise katse puhul on ühe lapsega tehtud katse vastustes kolm minevikuvormi ning ainult üks

[4] Sageli on lapsel kergem täiendada või parandada just mõne mängulooma vm kolmanda tegelase lauseid. COST A33 projekti „Crosslinguistically Robust Stages of Language Development" raames eesti lastega läbi viidud aspekti omandamise katses kasutati loomistesti treeninguks mängu, kus mängukaru kohta öeldi, et ta ei oska kunagi oma lauseid lõpetada ja et laps peaks teda aitama.

neist õiges ehk minevikulises kontekstis. Seetõttu peabki arvestama, et mida lühem on katse, seda rohkem peab olema katseisikuid. Hea oleks, kui üks tingimus (situatsioon ja selles kasutatud grammatiline vorm) esineks katses vähemalt kaks korda, näiteks kindlas käändes objekt kahe eri verbiga – see annaks võimaluse võrrelda sama tingimuse tulemusi (nt verbiti). Eesti keele objekti käändevahelduse omandamise katsest on selgunud (vt Argus 2009), et täpselt sama tingimuse korral võib katseüksusest arusaamist mõjutada ka objektnoomeni morfofonoloogiline struktuur ning seega peaks ühes katses olema vähemalt kaks sama struktuuriga objektnoomenit. Samas tuleb arvestada, et *pesa*-tüüpi noomenite (nt *maja, kala*) puhul ei ole genitiivi ja partitiivivormide vahel võimalik vahet teha ning seega ei saa katses neid sõnu, õigemini nende sõnadega tähistatavaid esemeid-rekvisiite kasutada.

Kindla vastamismustri (nt *ei-jah-ei-jah-ei-jah* jne) kasutamise vältimiseks tuleb katseüksused paigutada nii, et sellist kindlat vastuste mustrit ei tekiks. Lisaks on hea kasutada nn segajaid või täiteid (küsimusi), millele laps peaks vastama hoopis teisiti kui katseküsimustele, näiteks *ei-* või *jah*-vastust nõudvate küsimuste vahele tuleks paigutada *milline-* või *kus*-küsimusi. Selline varieerimine aitab katses vältida täpselt ühesuguste vastuste andmist[5].

On üldlevinud tava, et katse algusesse paigutatakse mõned treeningüksused, milles katse läbiviija võib lapse vastuseid kommenteerida ja parandada. Sageli kasutatakse treeningüksusi katseülesande selgitamiseks, vahel ka selleks, et teha

[5] Ajakategooria omandamise eelkatses (koostanud Bart Hollebrandse, katse eesti lastega läbi viinud Reili Argus ja Sirli Parm) leidus hulk lapsi, kes kippusid kasutama kõikide katseüksuste puhul ainult üht ajavormi; objekti omandamise katses (Argus 2009) leidus lapsi, kes vastasid kõikidele küsimustele jaatavalt.

kindlaks, kas laps tuleb üldse katsega toime. Näiteks kui normaalse keelelise arenguga lastele mõeldud katses ei vasta laps ühelegi treeningküsimusele õigesti, siis temaga katset ei jätkata. Lisaks võib treeningüksuste abil teha kindlaks, kas lapsel on olemas katse sooritamiseks vajalik taustateadmine (näiteks kas ta suudab pildil olevad tegelased või esemed ära tunda). Treeningüksuste hulk peaks olema optimaalne: liiga palju treeningüksusi venitab katse liiga pikaks. Kui eesmärk ei ole kontrollida, kuidas lapsed katse käigus (ehk treeningüksustega) mingi keelestruktuuri omandavad, siis on otstarbekas kasutada nii vähe treeningüksusi kui võimalik. Kolme tingimusega katse puhul oleks optimaalne treeningüksuste arv kolm.

Katse kestuse määramisel tuleb arvestada katses osalevate laste vanusega. Peab silmas pidama, et ka rühma noorim katsealune suudaks katse tüdinemata läbi teha. Näiteks nelja-aastastel lastel on raske 15 minutit järjest keskenduda, samas kuueaastase jaoks ei ole see sugugi raske. Kui katse on ülesehituselt vaheldusrikas ja kujunduselt atraktiivne, ei tüdine lapsed nii ruttu kui väga ühetaolise (samade tegelaste ja objektidega situatsioonid) katse puhul. Teisalt on aga oluline, et liigne atraktiivsus ei hakkaks katse keelelist poolt ehk ootuspäraste vastuste andmist segama ega sunniks last keskenduma katse seisukohast ebaolulisele, näiteks tegelaste näoilmetele vm. Lapse koostöövalmiduse tõstmiseks ja huvi hoidmiseks tasub kergemad küsimused paigutada võimalusel katse algusesse ja lõppu – nii ei teki lapsel katse alguses tunnet, et ta ei saa hakkama, ning katse lõpeb tundega, et tegemist ei olnudki millegi keerulise või raskega.

3. Eesti keele aspektilisuse omandamise uurimiseks kasutatud kahe erineva ülesehitusega katse tulemuste võrdlus

Järgnevalt on võrreldud ühe ja sama keelestruktuuri omandamise uurimiseks koostatud kahe erineva ülesehitusega katse[6] tulemusi, kasutades selleks vanuselt ja soolt kattuvat valikrühma. Rühm on vastavusse seatud nii, et mõlemas viiest lapsest koosnevas rühmas on laste keskmine vanus sama – 6 aastat 2 kuud. Poisse ja tüdrukuid on rühmades sama palju (3 poissi ja 2 tüdrukut).

Katsete üks eesmärk oli koguda andmeid selle kohta, millal omandab normaalse keelelise arenguga eesti laps aspektilisuse ehk objekti käändevahelduse, et saadud andmeid ja kogemusi kasutades edaspidi välja töötada selline katse, millega oleks võimalik määrata, kas alakõnega eesti lastel on kõnealuses valdkonnas probleeme või mitte. Teine eesmärk oli välja selgitada, mis on eesti objekti käändevahelduse omandamise juures raske: kas objekti käändevalikul eksitakse enam perfektiivsetes või imperfektiivsetes situatsioonides? Selleks vaadeldi katses tähenduse (perfektiivse – imperfektiivse tegevuse) ja vormi (objekt genitiivis või partitiivis) seost. Katsetes varieeriti lõpetatud-lõpetamata situatsiooni, ajavormiks läbivalt minevik. Kõik tegevused olid duratiivsed, mis võimaldas mõlema aspekti kasutamist.

Mõlemad katsed koosnesid mõistmis- ja loomistestist. Esimeses katses järgnes loomistest mõistmistestile, teises olid loomistesti üksused paigutatud mõistmistesti üksustega vahel-

[6] Katsed on koostatud COST A33 projekti „Crosslinguistically Robust Stages of Language Development“ aja- ja aspektiuurimise töörühmas.

dumisi. Katses kasutatud verbid valiti lähtuvalt nende varasest omandamisest[7] ning võimalusest verbiga väljendatud tegevust videofilmis demonstreerida. Situatsioonid esitatati videofilmil ja keeleline materjal suuliselt, st et mõistmistestis esitati lapsele kordamööda ühe või teise situatsiooni tüübi kohta käivaid lauseid, mis laps pidi kas õigeks või valeks tunnistama. Loomistestis tuli lastel katse läbiviija lauseid jätkata. Mõlemas katses oli kasutusel n-ö kolmas tegelane, kelle lauseid laps hindas ja jätkas.

3.1. Esimese katse üldkirjeldus

Katse[8] on legendilt võistlus, kus kaks tegelast, punases pluusis tüdruk ja mustas pluusis tüdruk, sooritavad ühte ja sama tegevust, näiteks söövad küpsist või ehitavad legodest silda. Kui klipp lõpeb, on üks tegelane oma tegevuse ilmselgelt lõpetanud (nt küpsis on söödud või sild on valmis), teine aga jätkab tegevust. Tegevuse lõpetanud osaline rõhutab lõpetatust sellega, et asetab käed lauale ja jääb ootama. Enne kui katse läbiviija lause lapsele hindamiseks esitab, paneb ta videoklipi seisma ning ekraanile jääb viimane kaader. Mõistmistestis öeldakse lapsele kordamööda perfektiivse ja imperfektiivse situatsiooni (ja vastava tegelase) kohta käivaid lauseid ning laps peab otsustama, kas lause on õige või ei (nt *Filmi lõpus ehitas punases pluusis tüdruk ehitas silla* või *... ehitas silda*). Loomistestis tuleb lapsel jätkata lauseid, nagu *Filmi lõpus sõi mustas pluusis tüdruk ...*, kasutades objekti õiges käändes. Mõistmistestis varieeriti testüksuseid vastavalt aspektisituatsiooni tüü-

[7] Varase omandamise all mõeldakse seda, kui vastav verb esineb lapse spontaanse kõne lindistustes enne kolmeaastaseks saamist.

[8] Kõnealust katset on kasutatud eesti keele objekti käändevahelduse alase kirjutise (Argus 2009) alusmaterjalina. Katse on koostanud Reili Argus, Angeliek van Hout ja Isabel Garcia del Real.

bile ja kasutatud käändevormile: perfektiivne situatsioon – testlauses genitiivivorm, imperfektiivne situatsioon – testlauses partitiivivorm, perfektiivne situatsioon – testlauses partitiivivorm ja imperfektiivne situatsioon – testlauses genitiivivorm. Loomistestis kasutati kordamööda perfektiivseid ja imperfektiivseid aspektisituatsioone.

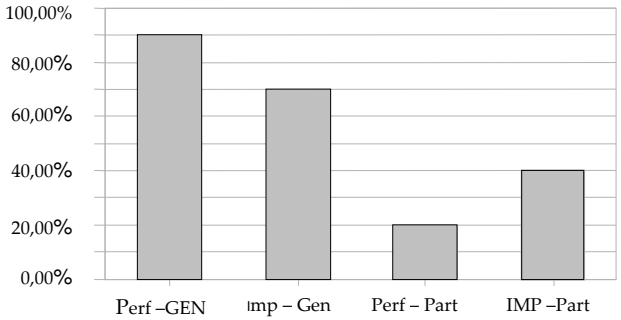
3.2. Teise katse üldkirjeldus

Katse[9] on legendilt mäng, kus tegelane võib ennast liigutada ainult siis, kui muusika mängib. Erinevalt esimesest katsest on siin kasutatud ühte tegelast – klouni, kes sooritab mingeid tegevusi (nt ehitab legodest autot) ja vahel saab muusika lõppedes tegevusega valmis, vahel jääb tegevus pooleli (nt legodest majal on muusika vaikides katus puudu). Lapsele esitatakse lauseid, mis käivad kord perfektiivse, kord imperfektiivse aspektisituatsiooni kohta (nt *Siis kui muusika mängis, tegi kloun autot* või *Siis kui muusika mängis, tegi kloun auto*). Erinevalt esimesest katsest, kus videoklipp lõppes siis, kui esimene tegelane tegevusega valmis sai, on teises katses kasutatud liikuvat kaamerat: kui muusika vait jääb ja tegelane kivikujuks tardub, liigub kaamera lähiplaanilt kaugplaanile ning tegevuse lõpetatuse või mittelõpetatuse, samuti objekti saab laps mõne sekundi vältel selgelt fikseerida.

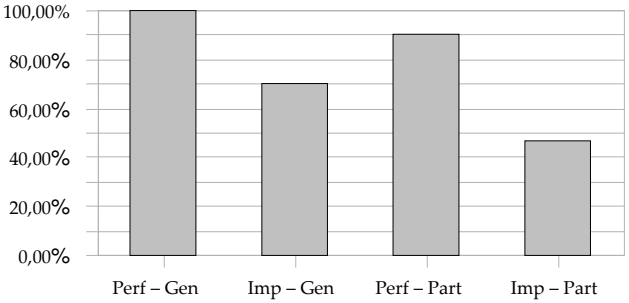
[9] Katse on koostanud Angeliek van Hout, Natalja Gagarina ja Margreet van Koert.

3.3. Kahe katse tulemuste võrdlus

Kahe katse mõistmistesti tulemustest annavad ülevaate joonised 1 ja 2. Tulemused on esitatud testis kasutatud tingimuste kaupa[10].



Joonis 1. Esimese katse mõistmistest: ootuspäraste vastuste hulk tingimuste kaupa



Joonis 2. Teise katse mõistmistest: ootuspäraste vastuste hulk tingimuste kaupa

[10] Joonisel kasutatud lühend Perf tähistab perfektiivset aspektisituatsiooni, Imp – imperfektiivset; Part tähistab testlauses kasutatud partitiivobjekti, Gen – genitiivobjekti.

Joonistest 1 ja 2 nähtub, et teises katses saadud tulemused on kõikides tingimustes esimese katse tulemustest paremad. Vahe ootuspäraste vastuste hulga vahel varieerub 8%st koguni 70%ni, olles kõige suurem just perfektiivses aspektisituatsioonis kasutatud partitiivikujulise objektiga testlausetes, kus lastel tuli lause valeks tunnistada. Mõlemas katses on kasutatud ajalisi piire. Vaatluspunkt on lausetes täpsustatud ajamäärustega (esimeses katses *filmi lõpus,* teises katses *siis kui muusika mängis*). Kuigi partitiivobjekti kasutus ei oleks ka perfektiivses aspektisituatsioonis päris vale (*siis kui muusika mängis, tegi kloun autot*), toob ajaline intervall, mitte tegevuse lõpphetk, nagu esimeseski katses, paremini esile just lõpetatuse tähenduse ning lapsed keskenduvad enam sellele, kas tegevus jõudis lõpule või ei. Kui võrrelda jooniseid 1 ja 2, siis on ilmne, et genitiivobjekti kasutamine on kuueaastaste eesti laste jaoks mõlemas katses selgem kui partitiivobjekti kasutamine.

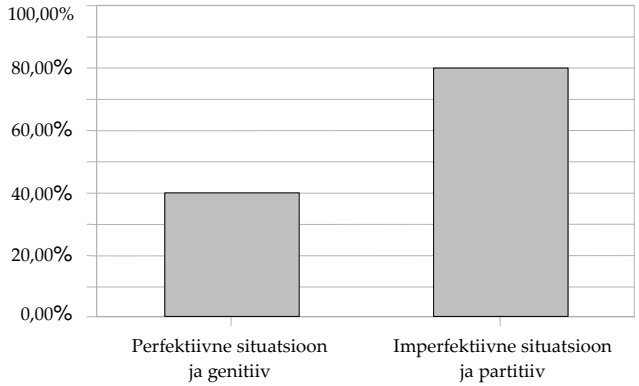
Huvipakkuv on asjaolu, et kahe *jah*-vastust nõudva tingimuse puhul (perfektiivne aspektisituatsioon ja genitiivobjekt ning imperfektiivne aspektisituatsioon ja partitiivobjekt) oli ootuspäraste vastuste hulk üsna erinev (vt joonised 1 ja 2). Kui perfektiivses situatsioonis kasutatud genitiivobjekti tunnistasid lapsed esimeses katses enamasti õigeks ja teises eranditult õigeks, siis imperfektiivse situatsiooni ja partitiivobjekti korral jääb õigeid vastuseid ka teises katses ikka alla 50%. Tekib küsimus, miks ei hinda lapsed lõpetamata tegevuse partitiivobjekti kastamist õigeks. On ju partitiiv see kääne, mille eesti lapsed varakult omandavad (Argus 2009).

Testüksuste ükshaaval vaatlemisest selgub, et eri verbidega testlausete õigete vastuste hulk mõnevõrra erineb. Näiteks on pildi tahvlilt *kustutamise* kohta õigeid vastuseid pisut rohkem kui auto *tegemise* kohta käivatel lausetel. Kõige väiksema hulga õigete vastustega paistis silma testüksus, kus kloun avas

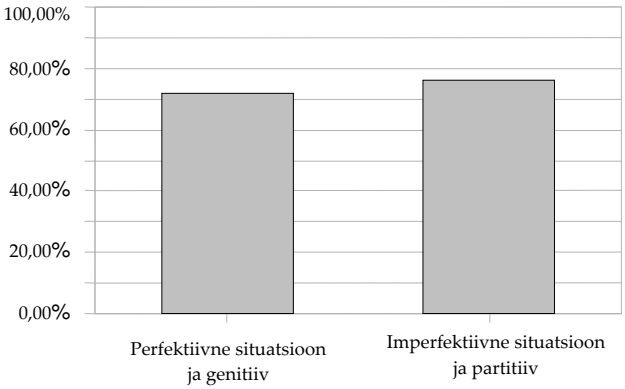
teekannu (vältevahelduslik objektnoomen). Samas ei paista objekti astmevahelduslikkus üldisi testitulemusi oluliselt mõjutavat, sest testüksuses, kus esines astmevahelduseta objektnoomen (*auto*), oli võrreldes enamiku astmevahelduslikku objektnoomeniga (nt *rong*, *sild*) testüksustest õigeid vastuseid vähem.

Ühe olulise tegurina tuleb siinkohal mainida hoopis intonatsiooni (vt Argus 2009). Nimelt sunnib objektnoomeni rõhutamine testlauses (*kloun sõi kooki* – st et *sõi just nimelt kooki*) lapsi keskenduma objektile, mitte tegevusele või selle lõpetatusele, mis tingib partitiivobjektiga lause õigeks tunnistamise. Seega tuleb tõdeda, et kummaski katses pole suudetud mõtestada ja kontrollida kõiki tulemusi mõjutavaid tingimusi – intonatsiooni mõju testlausete tõlgendamisele vajab edasist uurimist ja ilmselt ka eellindistatud testlauseid, mille puhul iga lause intonatsioon on kõikide katsealuste jaoks ühesugune.

Kahe katse loomistesti tulemustest annavad ülevaate joonised 3 ja 4. Tulemused on esitatud testis kasutatud tingimuste kaupa.



Joonis 3. Esimese katse loomistest: ootuspäraste vastuste hulk tingimuste kaupa



Joonis 4. Teise katse loomistest: ootuspäraste vastuste hulk tingimuste kaupa

Sarnaselt mõistmistestiga on ka loomistesti puhul teise katse tulemused paremad kui esimese omad: õigete vastuste hulk jääb ülespoole juhusliku valiku[11] piiri, eriti perfektiivsete aspektisituatsioonide puhul. Võib oletada, et paremad tulemused on tingitud sellest, et perfektiivsus oli teises katses paremini esile toodud. Videoklipi kaameratöö, kui muusika lõppedes liigub kaamera tegelaselt ja objektilt pisut kaugemale, andes lapsele n-ö fikseerimise pausi, mille jooksul tal on võimalik veenduda, et tegevus sai lõpetatud.

Veaanalüüs näitab, et teises katses keskendusid lapsed tegevusele ja objektile, mitte kahe tegelase vahelisele võistlusmomendile. Samuti oli vähem vastuseid, milles objekti polnud kasutatud (nt *Tegi kiiremini* või *Sai varem valmis*). Teises katses

[11] Tavaliselt loetakse õigete vastuste protsentuaalse jaotumise puhul juhuslikuks tasemeks vahemikku 40-60%; omandamise alguseks on enamasti peetud vanust, kus õigete vastuste protsent on üle 70% (Weist jt 1991: 78).

esines perfektiivsetes ja imperfektiivsetes aspektisituatsioonides erinevat tüüpi vigu: mõnel korral ei olnud perfektiivsuse kirjeldamiseks kasutatud objektnoomeni eeldatavat vormi, vaid vahendit märkivaid sõnu komitatiivis, nt *ehitas legodest*, *tegi klotsidest*; imperfektiivsuse edastamisel eelistati aga kasutada valedeks loetud genitiivobjektiga testüksusi (*tegi silla*, kuigi videoklipis oli näha, et sild ei saanud valmis).

4. Psühholingvistiliste katsete koostamise ja läbiviimisega seotud probleemidest ning nende vältimisest

Keeleomandamise uurimiseks kasutatavate psühholingvistiliste katsete väljatöötamisel tekib üldjoontes kolme erinevat laadi, kuid samas omavahel tihedalt seotud probleeme, mis puudutavad katse läbiviimisega seonduvat, katse ülesehitust ja konkreetse keele spetsiifikat.

Katse läbiviimisega seotud probleemideks võib pidada katse ajalist kestust ja selle optimaalsust, katse atraktiivsust, katsesituatsiooni loomulikkust ja keelelise raamistiku valikut (nt vajab läbimõtlemist, kuidas vältida sihtkeelendite n-ö ettesöötmist). Kirjeldatud katsete puhul ilmnes, et tähtsusetu ei ole ka see, millises järjekorras lastakse lastel mõistmis- ja loomistesti teha. Üldiselt on leitud[12], et mõistmistesti tulemused on loomistesti tulemustest paremad. Samas on saadud tulemusi, kus mõistmis- ja loomistesti tulemused ei erinenud,

[12] Arutelu rahvusvahelise projekti COST A33 „Crosslinguistically Robust Stages of Language Development" töökoosolekul 2009. aasta 14. jaanuaril Berliinis.

kuid sel juhul oli loomistest läbitud enne mõistmistesti.[13] Siinse uurimuse aluseks olevas esimeses katses, kus mõistmistest eelnes loomistestile, oli loomistestis mõnevõrra madalam õigete vastuste protsent kui mõistmistestis. Teises katses olid aga loomis- ja mõistmistesti üksused paigutatud vaheldusmisi ning loomistesti tulemused ei olnud mõistmistesti tulemustest sugugi halvemad. Järelikult on loomistesti üksuste vaheldumine mõistmistesti üksustega metoodiliselt igati õigustatud.

Teine probleemiring tuleneb katsega kontrollitava ehk omandatava keele struktuuri iseärasustest. Ajakategooria omandamise katses osutub problemaatiliseks tulevikusituatsiooniga seonduv. Kuna tulevik ei ole eesti keeles grammatiliseks kategooriaks kujunenud, siis ei saa lugeda valeks olevikuvormide kasutamist tulevikku suunatud tegevuse puhul. Teisalt, kuna tuleviku grammatilist kategooriat eesti keeles ei ole, siis on sedalaadi katse eriti informatiivne mitte ainult keeleomandamise kontekstis: ääretult huvitav oleks teada saada, milliseid keelevahendeid lapsed ja täiskasvanud tulevikulisusele viitamiseks üldse kasutavad.

Konkreetse keele struktuuri spetsiifika on tihedasti seotud testüksuste mitmekesisuse ja sedakaudu ka katse ülesehituse ja kestusega. Eesti keele eripärast tulenevalt on aspektilisuse, aga ka näiteks käändevormide omandamise alastes katsetes kasutatavate noomenite puhul vaja arvestada nende morfofonoloogilise eripära ja osa käändevormide homonüümsusega. Kuna eesti keeles on nii astmevahelduslikke kui ka astmevahelduseta noomeneid, siis tuleb katses, kus kontrollitakse mingi käändevormi omandamist, kasutada mõlemaid, lisaks ka laadi- ja vältevahelduslikke noomeneid. Paljude

[13] Isiklik vestlus Dagmar Bittneriga rahvusvahelise projekti COST A33 „Crosslinguistically Robust Stages of Language Development" töökoosolekul 2009. aasta 15. jaanuaril Berliinis.

indoeuroopa keelte puhul kasutatakse väikestele lastele mõeldud keelekatsetes ainult reeglipäraseid verbe. Eesti keele killustatud muutsüsteemi silmas pidades ei ole reeglipäraselt ja ebareeglipäraselt muutuvate verbide vahele võimalik nii selget piiri tõmmata, samas on oluline, et näiteks aspektilisuse omandamist kontrollivas katses oleks lihtverbide kõrval kasutataud ka erinevaid perfektiivsuspartikleid sisaldavaid ühendverbe (nt ***ära*** *sööma,* ***valmis*** *tegema*). Seega sõltub varieeritavatate katsetingimuste hulk ning katse üldine pikkus konkreetse keele struktuurist.

Eesti keele omandamise uurimise esimeste katsete koostamisest ja rakendamise kogemuste põhjal võib kokkuvõtvalt väita, et keeleomandamiskatse väljatöötamine on ääretult keerukas ja aeganõudev ülesanne. Arvestada tuleb paljude korralduslike, ülesehituslike ja keeleomaste teguritega. Katse ülesehitust tasub aga alustada just keeleomastest teguritest lähtuvalt. Tegevuse aspektiähenduste erinevuse omandamise kahe katse tulemuste kohatine 70%line erinevus osutab sellele, kuivõrd olulist rolli mängib katse ülesehitus. Teise katse ülesehituslikud muudatused, näiteks kaamera liikumine objektilt suurele plaanile ja sellega kaasnev paus kahe testüksuse vahel, samuti loomistesti üksuste paigutamine mõistmistesti üksustega vaheldumisi parandasid oluliselt testi tulemusi. Võib väita, et katsete koostamise ja läbiviimise metoodika täiustamine katset võimalikult suure kontrollgrupiga läbi viies ja võimalikult paljusid tegureid arvesse võttes on täiesti võimalik.

Kirjandus

Argus, Reili 2009. Psühholingvistiline katse eesti keele objekti käändevahelduse omandamise uurimise meetodina. – Emakeele Seltsi aastaraamat 54. Tallinn: Teaduste Akadeemia Kirjastus, 22–43.

Crain, Stephen & Thornton, Rosalind 1998. Investigations in Universal Grammar. – A Guide to Experiments in the Acquisition of Syntax and Semantics. Cambridge, MA: MIT Press.

Ehala, Martin 2007. Change in progress in the Estonian object marking system. – Paper presented at the School on Grammaticalisation and Typology, Rakvere, 22.03.2007. http://www.fl.ut.ee/orb.aw/class=file/action=preview/id=228558/Microsoft+PowerPoint+-+Estonian+object+marking_Ehala.pdf (20.02.2009).

Leonard, Laurence B. & Deevy, Patricia & Miller, Carol A. & Charest, Monique & Kurt, Robert & Rauf, Leila 2003. The use of grammatical morphemes reflecting aspect and modality by children with specific language impairment. – Journal of Child Language 30, 769–795.

Kosinski, Robert & Cummings, John 1999. The scientific method: An introduction using reaction time. – Tested studies for laboratory teaching 20 / Ed. by S. J. Karcher. Proceedings of the 20th Workshop/Conference of the Association for Biology Laboratory Education (ABLE), 63–84.

Welford, Alan Traviss 1980. Choice reaction time: Basic concepts. – Reaction Times Ed. by A. T. Welford. Academic Press, New York, 73–128.

Vinnitskaya, Inna & Wexler, Kenneth 2001. The role of pragmatics in the development of Russian aspect. – First Language 21, 143–186.

Weist jt 1991 = Weist, Richard & Wysocka, Hanna & Lyytinen, Paula 1991. A cross-linguistic perspective on the development of temporal systems. – Journal of Child Language 18, 67–92.

Using experiments for the Study on Estonian Language Acquisition

Reili Argus

Summary

Psycholinguistic experiments are not widely used in the research of the acquisition of Estonian. The overview of main types of psycholinguistic experiments used in the research of first language acquisition is presented in the first part of the article, comments relying on personal experience of the author are added to all types of experiments. The second part of the article consists of the description, analysis and results of two psycholinguistic experiments designed for the research of the acquisition of Estonian aspect. Problems concerning the design and the running of the experiment, are indicated, also some important details the researcher and the designer of the experiment must take into the consideration are pointed out. The amount of correct answers of children in the same condition in two different tests can vary from 8 to 70%. Main results of two different experiments of the acquisition of aspect indicate that the design and the structure of the experiment are extremely important.

Keywords: acquisition of Estonian, child language, designing of a psycholinguistic experiment, the structure of an experiment

Autor

PhD Reili Argus, Tallinna Ülikooli eesti keele ja kultuuri instituudi dotsent, lastekeeleuurija. COST A33 projekti „Cross-linguistically Robust Stages of Language Development" aja- ja aspektiuurimise töörühma liige, reili.argus@tlu.ee

EESTIKEELSES TEKSTILOOMES EELISTATUD KONSTRUKTSIOONID JA KÄÄNDEVORMID

Pille Eslon

Ülevaade

Artiklis võrreldakse käändevormide kasutuseelistusi eesti ajakirjanduskeeles ja eesti õppijakeeles (vahekeeles).[1] Kahte korpusainest on analüüsitud statistikal põhinevate programmide ja eesti keele tarkvara abil. Seejärel on võrreldud mõlema korpuse tekstiloomes kaks ja enam korda esinenud konstruktsioone, milles on võimalik kasutada ainult kindlaid käändevorme. Saadud tulemused kajastavad mitte üksnes eesti ajakirjandus- ja õppijakeele käändekasutuseelistusi, vaid ka nende kahe keelevariandi diskursuserinevusi. Uurimuse aluseks on korpusainestik, millel rakendatakse korpusest tulenevat analüüsisuunda.

Võtmesõnad: korpuslingvistika, korpusest tulenev võrdlev analüüs, käändevormide kasutamine, eesti keel

---

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

1. Lähtepunkt

Tänapäeva keeleteaduses on korpuste kasutamine uurimuste tarbeks tavapärane. Samas on korpust kui allikmaterjali kogu mõistetud erinevalt: uurija enese jaoks kitsal eesmärgil koostatud näiteainesest miljonisõnaliste elektrooniliste korpusteni, millel võib olla oma märgendussüsteem ja kasutajaliides. Erinev on olnud ka arusaam sellest, kuidas korpusi koostada ning uurida, milliseid analüüsimeetodeid ja -vahendeid rakendada, et leida vastuseid tõstatatud küsimustele, tõestada hüpoteese või välja selgitada keelendite funktsionaalne potentsiaal: vt nt inglise keele *if*-konstruktsiooni uurimus Gabrielatos 2007, soome keele muutkondade produktiivsuse indeksi määramine Nikolajev 2007, eesti keele *tulema*-verbide sageduse muutumine kindla kõneviisi ainsuse kolmanda pöörde vormis Kilgi 2006. Niisugust uurimissuunda on korpuslingvistikas enamasti nimetatud korpuspõhiseks keeleanalüüsiks (ingl k *corpus-based language analysis*), mille sünonüümina on Paul Rayson kasutanud mõistet *hypothesis-driven* ´hüpoteesist tulenev´ (vt Rayson 2002: 1) ning mida Geoffrey Leech on pidanud orienteerituseks teooriale või lingvistilisele koolkonnale (*theory-oriented paradigm characteristic of some other schools of linguistics*), vt Leech 2004: 61. Töö alguses tõstatatakse uurimisküsimus, luuakse vastavalt märgendatud korpus, mida analüüsitakse (pool)automaatselt ja / või käsitsi. Täiendavate uurimismeetoditena rakendatakse korpusainese kvantitatiivset ja kvalitatiivset analüüsi, mille tulemusi tõlgendatakse lingvistiliselt (vt Rayson 2002: 13; Granger 2003). Korpuspõhine keeleuurimine pole välistanud seda, et korpusi kasutatakse jätkuvalt allikmaterjalina oma teoreetiliste seisukohtade illustreerimiseks (*corpus-illustrated language research*).

Teine tänapäevane võimalus keelt uurida on korpusest või korpusainestikust tulenev uurimissuund (*corpus-driven research*

või *data-driven research*), mille puhul ei käsitleta korpust pelgalt oma teoreetilistele seisukohtadele kinnituse leidmise allikana. Leechi järgi on see keeleuurimise teine paradigma, mida ta on nimetanud andmetest tulenevaks (*data-driven paradigm characteristic of corpus linguistics*), vt Leech 2004: 61. Kuna korpuse keeleaines kajastab keelekasutusele iseloomulikke nähtusi, siis on oluline leida moodus, kuidas need nähtused esile tuleksid. Võimaluse selleks peaks andma korpusainese uurimine mittelingvistiliste vahenditega, rakendades erinevaid statistikal põhinevaid analüüsimeetodeid (nt keeletarkvara, statistiline andmetöötlus), mis välistavad uurija subjektiivsed valikud ning toovad esile uusi aspekte keeles toimivate seoste ja protsesside kohta, nt varjatult kulgevad keelemuutused ja ootamatud arengutendentsid. Tähelepanu keskmesse tõusevad seosed lingvistiliste struktuuride ja nende kasutussageduse vahel. Sel moel tuleb esile nii lingvistiline kui ka mittelingvistiline teave, mida korpusstatistika uurijale pakub. Just need andmed peaksid andma impulsi teoreetilist laadi mõttekäikudele. Uuritakse, kuidas on omavahel põimunud keelekasutus (ingl k *use of language*), sünkroonne varieerumine (*synchronic variation*) ja diakroonilised keelemuutused (*diachronic change*), vt Leech 2004: 77–78. Selle suuna eelis peitub võimaluses tõstatada hüpoteese ning luua teooriaid uurimisobjektist tulenevalt, mitte etteantud teoreetilistest seisukohtadest lähtudes, mis omakorda lubab viia keelekirjelduse vastavusse kirjeldatava objekti olemusega ja vältida lingvisti subjektiivsete eelistuste ning arusaamade võimalikku eksitavat mõju uurimistulemustele (vt Tognini Bonelli 2002). Korpusest tuleneva uurimise aluseks võib olla kas märgendatud korpus või tekstiarhiiv, mida analüüsitakse kvantitatiivselt. Saadud tulemused näitavad, mida antud korpuse põhjal tasuks üldse uurida, mis on keelekasutuses sage, mis mitte (vt ka Rayson 2002: 14), missugused lingvistilised mustrid on

keelekasutuses olulised, kuidas neis on põimunud leksika ja grammatika. Analüüsis ei lähtuta nt pelgalt sõnavarast või grammatilistest vormidest, sõnade süntaktilistest funktsioonidest, rollidest lauses jne. Oluline on keeleline väljund, mis kajastab semantika, grammatika ja pragmaatika koosmõju lausungis, laiemalt tekstis, diskursuses[2].

Kolmas võimalus on lähtuda mõlemast kirjeldatud uurimissuunast, varieerides neid vastavalt uurimiseesmärgile ja analüüsitava materjali iseloomule. Leitakse, et ei korpuspõhine ega ka korpusest tulenev analüüs pole universaalne uurimismeetod, millel on võrdsed võimalused ühelt poolt mistahes keelendite uurimisel ja teisalt diskursusanalüüsis, eriti kui diskursust mõistetakse verbaalsest tekstist avaramalt (vt Baker 2006: 17–21). Samuti arvatakse, et uurimustes, mille aluseks on tasakaalustatud selgetel printsiipidel rajanevad korpused, kus tekstid on rangelt valitud, esindavad kindlat tüüpi diskursuseid jne, on erinevaid analüüsimeetodeid mõttekas rakendada integreeritult (vt Behrens 2008: XXVII; Orpin 2005: 38–39). Samuti soovitatakse kombineerida erinevates teadusharudes rakendatavaid meetodeid – statistilisi, lingvistilisi, psühholingvistilisi, sotsiolingvistilisi jt (vt Taylor 2008: 183). Hindamaks seda, missuguseid lingvistilisi mustreid leidub erinevates tekstiliikides, millise diskursusega nende mustrite kasutamine on seotud, peetakse korpuslingvistikas vajalikuks võrrel-

[2] See korpuslingvistiline seisukoht haakub välja teooria ning funktsionaalgrammatilise suunaga keeleteaduses. Alates Praha lingvistilisest koolkonnast ja lõpetades näiteks Peterburi-Moskva keeletüpoloogide (V. S. Hrakovski, V. A. Plungjan jt) ning grammatikateooria uurijatega (A. V. Bondarko, J. A. Pupõnin jt) on olnud läbivaks ideeks keelendite ja kategooriate koosmõju lausungis, kus avaldub nende funktsionaalne potentsiaal, mis on omane eelkõige leksikaalgrammatilisele perifeeriale.

da erinevaid terviktekste ja testida saadud tulemusi suuremahuliste kirjakeele korpuste peal (vt nt Stubbs 1996: 126–154). Niisugust suunda on korpuslingvistikas nimetatud korpuste abil teostatavaks analüüsiks (ingl k *corpus-assisted analysis*). Seda võimalust on kasutatud õpikutekstide ja muude keeleõppematerjalide sõna- ning vormikasutuse autentsuse hindamisel (vt nt Römer 2007), registrierinevuste esiletoomisel (nt Biber 2004) jne. Lisaks klassikalisele ükskeelsele korpusainesele on peetud vajalikuks sama nähtuse uurimisse kaasata ka teisi allikaid (veebimaterjalid, spetsiaalselt antud uurimiseesmärgil produtseeritud tekstid jne), erinevat tüüpi korpuseid (nt paralleel- ja tõlkekorpused).

Olles valinud mistahes uurimissuuna (korpuspõhine, korpusest tulenev, integreeritud lähenemine) ning analüüsimeetodid, tuleb neid rakendades olla süstemaatiline, püüda haarata keeleainest kogu selle rikkuses, olgugi et kitsaskohti esineb eelkõige suulise kõne korpustega. Selle metodoloogilise lähtekoha olulisust on rõhutanud ka Jarmo Harri Jantunen (2009), kui ta kirjutab vajadusest uurida keelt mitte ainult süsteemselt, vaid süstemaatiliselt korpusainest kasutades ning erinevaid uurimissuundi, samuti korpuste kvantitatiivse ja kvalitatiivse analüüsi võimalusi integreerides. Niisugune põhimõte on klassikaline: midagi ei saa välistada, kuid olles valiku langetanud, tuleb seda ka süstemaatiliselt rakendada.

Käesolevas artiklis esitatavas eesti keele käändekasutustendentside uurimuses on erinevate korpusainestike võrdlemisel lähtutud korpusest tulenevast analüüsisuunast, rakendatud erinevaid statistikal põhinevaid ning keelest sõltumatuid analüüsiprogramme ja eesti keele tarkvara. Aluseks on Douglas Biberi multidimensionaalse analüüsi idee (vt Biber 2004: 15, 17–18), mis haakub hästi artikli autori üldmetodoloogilise tõekspidamisega uurimistöö ja taksonoomiate loomise mitme-

mõõtmelisusest (vt Eslon 2006). Biberi mõte seisneb selles, et uurimistöös tuleb kasutada erinevat statistikat, eriti faktor- ja klasteranalüüsi, kuna need annavad hea võimaluse registreerida keelevariatiivsusega seotud nähtusi (nt registri varieerumist inglise keele suulises ja kirjalikus esituslaadis) ning võrrelda erinevate keelekasutusvariantide lingvistilisi mustreid. Kui multidimensionaalse analüüsi tulemused viitavad samadele nähtustele, siis ei ole põhjust uurimistulemuste objektiivsuses kahelda.

Uurimisobjekt on praegusaja eesti keelt kõige ehedamalt kajastav ajakirjanduskirjakeel[3] ja õppijakeel[4] – eesti keele kaks aktiivselt kasutatavat varianti. Korpusstatistika alusel on neis välja toodud neli sagedasemat nimisõna *inimene*, *elu*, *aeg* ja *sõna*[5], mille käändevormide kasutusmustreid kirjalikus tekstiloomes on uuritud statistikal põhineva WordSmith Tools 5,0 programmidega ning eesti keele süntaksianalüsaatori abil; saadud väljundit on töödeldud selleks spetsiaalselt programmeeritud makrode abil[6]. Neid meetodeid omavahel kombineerides ja erineva korpusainese suhtes rakendades õnnestub välja tuua konstruktsioonid, mida eesti ajakirjanduskeele ja õppijakeele tekstiloomes on esinenud kaks ja enam korda ning milles omakorda on võimalik kasutada vaid kindlaid käände-

---

[3] Eesti Keele Instituudi keelekorpus = EKI korpus, vt http://www.eki.ee/corpus/ (03.09.2008).

[4] Eesti vahekeele korpus = EVKK, vt http://evkk.tlu.ee (03.09.2008).

[5] Vt Eslon & Matsak 2009; sagedasemate nimisõnade väljatoomise protseduuri on kirjeldatud artiklis Eslon 2008: 33–35.

[6] Programmeerimistöö on teinud Erika Matsak projekti „VAKO: Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine (2008–2010)" raames. Valimi representatiivsuse ning andmete võrreldavuse põhjendust vt Eslon & Matsak 2009.

vorme[7]. Ühelt poolt tulevad nii ilmsiks loomuliku keelekasutuse ning eesti õppijakeele diskursuse sarnasused ja erinevused. Teisalt annab korpusest tulenev statistikal põhinev analüüs teavet konstruktsioonitüüpide ja grammatiliste vormide valikust, kuvades pildi sellest, mida kujutab endast loomulik keelekasutus, mida õppijakeel. Ilmnevatel erinevustel on nii teoreetiline kui ka praktiline väärtus: ühelt poolt näitab uurimus olulisi tendentse eesti keele käändegrammatika arengus, teisalt aga võimaldab välja tuua loomuliku eesti keele ja eesti õppijakeele mõningad diskursuserinevused, tekstiloomes aktiivselt kasutatavad konstruktsioonitüübid ja käändevormid. Eristatakse 1) mõlema keelevariandi stereotüüpseid ehk kõrgsagedasi konstruktsioone, nt ajakirjanduskeeles *on (käes) teoinimeste aeg*; *sõna otseses mõttes (tähenduses)* ja õppijakeeles *vabadus (vajadus) planeerida aega* jne; 2) ühe keelevariandi keskseid ehk ainult sellele keelevariandile tüüpilisi konstruktsioone, nt õppijakeelele iseloomulikud konstruktsioonid *igal inimesel on (peab olema)*; *sõnal võib olla* jne; 3) erinevusi samalaadsete konstruktsioonide sageduses ja leksikaalgrammatilises varieeruvuses, nt kvantorfraasi kasutamine ajakirjanduskeeles *ligi (veel, umbes) 200 (kolm, 800) inimest; hukkus kolm (kaheksa, kuus, viis, üksteist) inimest* jt ning õppijakeeles *ainult (umbes) kuus (sada) inimest; ning kaheksa (üksteist) inimest, kaheksa (kuus) inimest 32-st; üksteist inimest kirjutasid* jne. Neid konstruktsioone nimetatakse tavapärasteks.

Analüüsiüksus eraldatakse formaalselt ning piiratakse kolmest sõnast koosneva keeleüksusega, mida korpuse statistikas esineb antud kujul kaks ja enam korda. Need keeleüksused ei järgi konstruktsioonipõhist süntaksimudelit, mis tugineb verbi argumendistruktuurile või noomenisüntaksile (vt Goldberg

[7] Statistilise analüüsi on teinud Erika Matsak, vt Eslon & Matsak 2009.

1995). Ka pole need sõnakombinatsioonid otseselt seotud loogiliste struktuuridega, vaid on välja toodud statistiliselt. Käesolevas uurimuses on konstruktsioonide sageduse määraks võetud viis ja enam korda, millest piisab ülevaate saamiseks olulisematest tendentsidest eesti ajakirjandus- ja õppijakeele tekstiloomes eelistatud konstruktsioonidest ning käändevormidest. Seejuures ei jälgita, kas tegu on mingi kindlat tüüpi fraasi, moodustaja süntaktilise funktsiooni, lausekonstruktsiooni, vaba sõnaühendi, püsiühendi või idioomiga. Kolmest sõnast koosnevad keeleüksused ehk konstruktsioonid on statistika põhjal välja toodud sõnavormide kombinatsioonid. Samuti ei saa neid pidada mitmesõnalisteks väljenditeks (ingl k *multi-word expressions*), kuna sellel mõistel on kindel sisu. Näiteks Francesca Masini vaatleb mitmesõnaliste väljendite all ka partikkelverbe (Masini 2005: 145–146), Kadri Muischnek idioome, poolidiomaatilisi ja kollokatiivseid ühendeid (Muischnek 2006: 12–22, 26–27). Nadja Nesselhauf räägib sageduse alusel välja toodud süntaktilistest ja semantilistest terviküksustest (Nesselhauf 2005: 21).

Käesolevas uurimuses on statistilise analüüsi tulemusel leitud korpusainesest hulk samalaadseid korduvaid konstruktsioone, millest osa moodustavad suvalised tähenduseta, sisutühjad ja absurdsed sõnavormide, arvude, sidendite, isikunimede, morfeemide, tähemärkide jne kooslused (nt *25 0 inimest*, *ka vägi elus*, *kolmsada kolmesaja inimene*, *inimene inimest külas*, *d aeg lustama*, *120 sõna töö* jne). Need statistiliselt välja toodud üksused eraldati tähendust omavatest konstruktsioonidest (nt *sõna sekka öelda, ehk teiste sõnadega, tol ajal oli, sellel ajal oli, sel ajal tekkis* jne) ning jäeti analüüsist välja. Seetõttu kasutataksegi antud uurimuses mõistet *konstruktsioon*, mille all peetakse silmas erinevaid struktuure, mis on välja toodud statistiliselt, mis koosnevad kolmest keeleüksusest, millel on lauses oma kindel tähendus ja süntaktiline funktsioon, nt lausekonstrukt-

sioon *et inimesed on* alustab kõrvallauset – *Ants ütles mulle, et inimesed on talle palju liiga teinud*; *Et inimesed on enamasti unustanud, mis toimus viiskümmend aastat tagasi, siis on meie kohus ...* jne. Igal konstruktsioonil võib olla analüüsitavale keelevariandile iseloomulik leksikaalgrammatiline varieeruvus (nt ajakirjanduskeele stereotüüpses konstruktsioonis *sõna otseses mõttes (tähenduses) ~ selle sõna otseses (kõige <otsesemas mõttes, tähenduses>)*), mida nimetatakse antud konstruktsiooni kasutusmustriks. Viimase invariandiks on tekstiloomes kõige sagedamini kasutatud konstruktsioon – antud juhul *sõna otseses mõttes*. Samuti võivad konstruktsioonid olla omavahel seotud kui süntaktilised sünonüümid, nt *mõni aeg hiljem ~ mõni aeg pärast <mida> ~ mõne aja pärast* jne.

Niisiis peetakse käesolevas uurimuses oluliseks, missuguseid kolmest keeleüksusest koosnevaid statistilisi konstruktsioone eelistatakse tekstiloomes kasutada. See võimaldab konstruktsioone tõlgendada kasutuspõhiselt.

2. EKI tekstikorpuses ja EVKK tekstides eelistatud konstruktsioonid ning käändevormid

Konstruktsioonid eristuvad vastavalt sagedusele, tüübile ja leksikaalgrammatilisele varieeruvusele. Omaette jaotustena kirjeldatakse eesti ajakirjandus- ja õppijakeele stereotüüpseid, tavapäraseid ning keelevariandi keskseid ehk tüüpilisi konstruktsioone ja nende leksikaalgrammatilisi variante.

2.1. Stereotüüpsed konstruktsioonid ja nende leksikaalgrammatilised variandid

Stereotüüpsete konstruktsioonide all mõistetakse ühes või teises keelekasutusvariandis regulaarselt ilmnevaid konstrukt-

sioone ja nende leksikaalgrammatilisi variante, mida kasutatakse tavaliselt hinnangut väljendava markeri või tekstisidususvahendina. Ülisuure sageduse tõttu peetakse neid tekstiloomes enamasti parasiitväljenditeks, mida keelekorraldajad soovitavad vältida. Samas kasutatakse eesti ajakirjanduskeeles näiteks sõnade *aeg* ja *sõna* käändevorme kindlates stereotüüpsetes konstruktsioonides, mida on vaja läinud hinnangu edastamiseks või siis tekstisidususvahendina: *on (käes) teoinimeste aeg*; *sõna otseses mõttes (tähenduses)*; *selle sõna otseses (kõige <otsesemas mõttes, tähenduses>)*; *tema (oma) sõnade kohaselt (järgi)*; *tehnika viimane sõna ~ oma viimane sõna ~ oma viimase sõna*; *oma sõna öelda (ütlema)*.

Need kirjakeele stereotüübid ei ole omased õppijakeelele – seda iseloomustavad teistlaadsed leksikaalgrammatilised konstruktsioonid, mille ülisage korduvus korpusaineses viitab ühelt poolt keeleõppe teemakesksusele ja teisalt õppijakeele korpuse tekstide annoteerimisega seonduvatele asjaoludele.

Teemakesksus väljendub nt konstruktsioonis *vabadus (vajadus) planeerida aega*, mida on kasutatud stereotüüpsena seoses oma positiivse suhtumise väljendamisega aja planeerimise vajalikkusse jms; konstruktsiooni *noorkeraamika (kammkeraamika) inimesed sarnanesid* kasutamine stereotüüpsena on seletatav kultuurilooteemaliste esseede ja kontrolltööde rohkusega EVKKs. Neis tekstides on pikalt kirjutatud Eesti riigi ajaloost. Samal põhjusel kuuluvad stereotüüpsete hulka ka konstruktsioonid sõna *aeg* kindlate käändevormidega: 1) lausekonstruktsioonid *et (kuid, sest) sel ajal ~ sest tol ajal ~ ja samal ajal*; 2) verbikesksed lausekonstruktsioonid *tol (sellel) ajal tekkis (kasutati) ~ samal ajal läksid ~ praegusel (sel, sellel) ajal on ~ vabariigi (vene, *tsari) ajal oli ~ sõja ajal polnud ~ vabal (vabariigi) ajal teha*; 3) noomenikonstruktsioonid *samal ajal ajaloolised <sündmused> ~ tol ajal inimesed ~ sel ajal laulud*; 4) modaalne konstruktsioon *külmal ajal peab*.

Õppijakeele korpuse tekstide annoteerimisega seonduvatest asjaoludest johtuvalt on stereotüüpide hulka sattunud ka tööjuhised, mis EVKKs on tekstist eraldamata. See küsimus lahendatakse korpuse edasise arendamise käigus, kui luuakse uusi alamkorpusi, täiendatakse ja ühtlustatakse metainfot, vaadatakse veel kord läbi korpuses sisalduvad tekstid, parandatakse sisestamisapsud, kontrollitakse üle sisseskanneeritud tekstid, eraldatakse õppija kirjutatud tekstiosa pealkirjast ning tööjuhistest jms. Praegu on tööjuhised õppijakeele stereotüüpsete konstruktsioonide hulgas, milles on *sõna* käändevorme kasutatud: 1) imperatiivses lausekonstruktsioonis *pane sõna(d) õigesse* <vormi>; *moodustage etteantud sõnadega* ja 2) kvantorfraasis *umbes 80 (120, 160) sõna*. Iseenesest pole selles midagi taunitavat, et tööjuhised kuuluvad õppijakeele stereotüüpide alla, sest need konstruktsioonid on keeleõppeprotsessi lahutamatu osa.

2.2. Tavapärased konstruktsioonid ja nende leksikaalgrammatilised variandid

Võrreldavate keelevariantide tavapäraseid konstruktsioone iseloomustab samalaadsus, erinevusi ilmneb vaid konstruktsioonide sageduses ning leksikaalgrammatilises varieeruvuses. EKI tekstikorpuse ja EVKK tekstiloomes eelistatud tavapärastes konstruktsioonides on eesti keele neljateistkümnest käändest kasutatud vaid grammatilisi – nominatiivi, genitiivi ja partitiivi, kusjuures valdav käändevorm on nominatiiv. Järgnevalt kirjeldame neid konstruktsioone käänete kaupa.

2.2.1. Ainsuse ja mitmuse nominatiiv

Nii loomulikus keelekasutuses kui ka eesti õppijakeeles on nominatiivi sisaldavad konstruktsioonid olnud kõige sagedasemad ja kõige rikkalikuma leksikaalgrammatilise varieeru-

vusega. Lahknevused esinevad nii konstruktsiooni tüüpides kui ka nende leksikaalgrammatilistes variantides, mis annab alust rääkida eesti ajakirjandus- ja õppijakeele diskursuserinevustest sõnade *inimene, elu*, *aeg* ja *sõna* kasutamisel.

EKI tekstikorpus

Ainsuse nominatiivi vormi sisaldavad konstruktsioonid:
1) loogiline implikatsioon *kui inimene on ~ et inimene on*
2) eitav verbikeskne lausekonstruktsioon *inimene ei ole (ei saa)*
3) lausekonstruktsioon *on ilus (paras, viimane, õige) aeg*; *aeg on kallis ~ aeg on möödas (läbi, käes)*; *oli aeg mil*
4) ajatähenduslik kvantorfraas *mõni aeg hiljem ~ mõni aeg pärast <mida>*; *on (oli, olnud, ajanud) kogu aeg*
Mitmuse nominatiivi vormi sisaldavad konstruktsioonid:
1) lausekonstruktsioon *need inimesed kes (kellele)*
2) jaatav (eitav) verbikeskne lausekonstruktsioon *et inimesed on (ei ole) ~ et (ka) need inimesed*; *need inimesed on*; *tema (ta) elu on*

EVKK

Ainsuse nominatiivi vormi sisaldavad konstruktsioonid:
1) loogiline implikatsioon *kui inimene on* (variandid: *kui inimene tahab, kogeb, elab, suhtleb*) *~ et inimene peab ~ ja kui (et kui, sest kui) inimene*
2) verbikeskne, enamasti modaalne lausekonstruktsioon *inimene peab olema (maksma, teadma) ~ inimene võib töötada*; *haritud (iga) inimene peab (saab, võib, tahab)*; *inimese elu on (võib, sõltub, erineb)*; *elu sõltub elukohast*, <meie> *elu erineb Ukraina* <elust>; *elu on väga*
3) mittemodaalne lausekonstruktsioon *see aeg on ~ mõni aeg oli*; *muutub (on) kogu aeg*; *on aeg kus*; *esimene sõna mis*; *lendab aeg väga*
4) verbikeskne eitav lausekonstruktsioon *inimene ei saa (ei ole)*

5) noomenikonstruktsioon *sõna eesti kultuur, väga levinud sõna ~ nö esimene sõna*

Mitmuse nominatiivi vormi sisaldavad konstruktsioonid:
1) jaatav (eitav) verbikeskne lausekonstruktsioon *kõik (need) inimesed on (hoidsid, tahavad)*; *inimesed kes on (ei ole)*

2.2.2. Ainsuse ja mitmuse genitiiv

EKI tekstikorpuses esineb suhteliselt harva genitiivi sisaldavaid konstruktsioone. Ajakirjanduskeeles tuli esile vaid ainsuse genitiivi sisaldav kvantorfraas. Ka EVKKs esines seda käänet teiste grammatiliste käänete vormidega võrreldes vähem. Õppijakeeles kasutati ainsuse genitiivi kahte liiki lausekonstruktsioonides, ajatähenduslikus kvantorfraasis ja noomenikonstruktsioonis. Mitmuse vormi kasutati küll sagedamini, kuid peamiselt ülivõrret sisaldavas noomenikonstruktsioonis ja lausekonstruktsioonis *on läbi aegade ~ läbi aegade on*; *on kõigi aegade*.

EKI tekstikorpus

Ainsuse genitiivi sisaldavad konstruktsioonid:
1) kvantorfraas *üle 200 (5000, 300, tuhande, saja) inimese*

EVKK

Ainsuse genitiivi sisaldavad konstruktsioonid:
1) jaatav (eitav) verbikeskne lausekonstruktsioon *mõjutab (ei mõjuta) inimese elu ~ inimese elu mõjutab*
2) lausekonstruktsioon *ja vanemate elu ~ et noorte elu ~ et inimese elu*
3) ajatähenduslik kvantorfraas *mõne aja pärast <tuli (oli)>*
4) noomenikonstruktsioon *inimese elu kõige ~ ?inimese elu haridus ~ *inimese elu tema* (?teema) ~ *inimese elu ma* (?maal)

Mitmuse genitiivi sisaldavad konstruktsioonid:
1) verbikeskne lausekonstruktsioon *kahjustab inimeste tervist*
2) noomenikonstruktsioon *kõigi aegade suurim (parim, esimene, teine)*; *ESTO-d läbi aegade*; *Eesti kõigi aegade*; *hümni sõnade kirjutaja (autor)*; *Eesti hümni sõnade*
3) lausekonstruktsioon *on läbi aegade ~ läbi aegade on*; *on kõigi aegade*; *sõnade autor oli*; *et hümni sõnade*

2.2.3. Ainsuse ja mitmuse partitiiv

EKI tekstikorpuse valimis on aktiivselt kasutatavatest konstruktsioonidest ainsuse partitiiviga eelistatud rikkaliku leksikaalse varieeruvusega kvantorfraasi. Mitmuse vormi esines sageli lausekonstruktsioonis *inimesi kellel on ~ inimesi kes ei; neid inimesi kes*; *on inimesi kes*. EVKKs on ainsuse genitiivi vormiga kasutatud erinevat tüüpi konstruktsioone rohkem kui EKI tekstikorpuses, ent valdavaks osutus samuti kvantorfraas. Ka mitmuse vorm esines samas konstruktsioonitüübis.

EKI tekstikorpus

Ainsuse partitiivi sisaldavad konstruktsioonid:
1) kvantorfraas *ligi (veel, umbes) 200 (kolm, 800) inimest; hukkus kolm (kaheksa, kuus, viis, üksteist) inimest; kuu aega tagasi (hiljem, enne); umbes (ligi, ja) kuu aega; juba pikka aega ~ väga pikka aega; oli (on) pikka aega ~ pikka aega on; tükk aega tagasi (pärast) ~ nädal aega tagasi; on veel aega ~ vajan veidi aega*

Mitmuse partitiivi sisaldavad konstruktsioonid:
1) lausekonstruktsioon *inimesi kellel on ~ inimesi kes ei; neid inimesi kes*; *on inimesi kes*

EVKK

Ainsuse partitiivi sisaldavad konstruktsioonid:
1) kvantorfraas *ainult (umbes) kuus (sada) inimest; ning kaheksa (üksteist) inimest, kaheksa (kuus) inimest 32-st; üksteist inimest kirjutasid*; *väga vähe aega ~ nii palju aega ~ liiga palju aega*
2) modaalne kvantorfraas **pidis (*pidab, pidada, pidama) mõnda aega*
3) verbikeskne lausekonstruktsioon *inimese elu mõjutab ~ mõjutavad inimese elu*; *kuulen sõna kultuur*, *ma kuulen sõna*, *kui kuulete sõna*, *on mõeldud sõna ~* <on> *mõeldud sõna neljakesi*
4) jaatav (eitav) verbikeskne lausekonstruktsioon *ei ole (ei olnud) aega*

Mitmuse partitiivi sisaldavad konstruktsioonid:
1) kvantorfraas *väga (nii) palju inimesi*

2.3. Eesti ajakirjandus- ja õppijakeelekesksed konstruktsioonid ning nende leksikaalgrammatilised variandid

Keelevariandi kesksed ehk tüüpilised konstruktsioonid on iseloomulikud vaid ühele võrreldavatest keelevariantidest. Kuigi nad kuuluvad tekstiloomes eelistatud konstruktsioonide alla, ei saa neid samastada stereotüüpsete konstruktsioonidega, kuna nad pole ülisagedased ega teatud määral tuhmunud tähendusega hinnangulisust edastavad või tekstisidususe eesmärgil kasutatavad markerid. Õppijakeeles kuuluvad siia need konstruktsioonid, milles on sageli kasutatud semantiliste käänete vorme: ainsuse adessiivi, inessiivi ja mitmuse komitatiivi.

Ainsuse adessiivi sisaldavad konstruktsioonid:
1) loogiline implikatsioon *kui inimesel on (ei ole)*
2) modaalne verbikeskne lausekonstruktsioon *igal inimesel on (peab olema)*; *sõnal võib olla*

Ainsuse inessiivi sisaldavad konstruktsioonid:
1) verbikeskne lausekonstruktsioon *minu elus on (oli) ~ meie elus on*; *elus on palju (väga)*

Mitmuse komitatiivi sisaldavad konstruktsioonid:
1) verbikeskne lausekonstruktsioon <minu> *jaoks seostub sõnadega*

Ajakirjanduskeele tekstiloomele polnud semantiliste käänete vorme sisaldavad konstruktsioonid sedavõrd omased, et oleksid mahtunud kasutuseelistuste tippu. Ajakirjanduskeelekeskseks osutus EKI tekstikorpuse valimi alusel hoopis nominatiivne substantiivkonstruktsioon (*inimene ja seadus ~ inimene ja loodus*; *haridus ja elu ~ kultuur ja elu*), mida kasutatakse ajalehele iseloomulikus funktsioonis – enamasti pealkirja või eraldi fraasina lugeja tähelepanu köitmiseks.

Õppija- ja ajakirjanduskeelekeksed konstruktsioonid ei ole omavahel üksüheses vastavuses, kuid seda ei saa hinnata ei õppijakeele puudusena ega loomuliku keelekasutuse rikkusena. Tegu on mõlemale eesti keele kasutusvariandile funktsionaalselt omaste konstruktsioonitüüpidega.

3. Kokkuvõte

EKI tekstikorpuse ja EVKK analüüsi alusel välja toodud tekstiloomes aktiivselt kasutatavad konstruktsioonid jagunesid kolme rühma: stereotüüpsed, tavapärased ja ajakirjandus- või õppijakeelekeksed ehk tüüpilised. Esimesse ja viimasesse rühma kuuluvad konstruktsioonid kajastavad mõlema keelevariandi eripära. Tavapäraste konstruktsioonide võrdlemine annab aga ettekujutuse loomuliku keelekasutuse ja õppijakeele universaalidest ning võimaldab välja tuua erinevusi konstruktsioonide leksikaalgrammatilises varieeruvuses ja kasutusmustrites.

3.1. Universaalsed nähtused

Selgus, et kõige universaalsem on loogilist implikatsiooni sisaldav konstruktsioon sõna *inimene* ainsuse nominatiivi vormiga: 1) ajakirjanduskeel – *kui inimene on ~ et inimene on*; 2) õppijakeel – *kui inimene on ~ et inimene peab*. Erinevus loomuliku keelekasutuse ja õppijakeele vahel seisnes vaid konstruktsiooni leksikaalses varieerumises olemis- ja kogemisverbidega, mis oli iseloomulik õppijakeele tekstiloomele: *kui inimene on / tahab / kogeb / elab / suhtleb*.

Teine universaalne konstruktsioon sõna *inimene* ainsuse nominatiivi vormiga on seotud eitusega: 1) ajakirjanduskeel – *inimene ei ole / ei saa*; 2) õppijakeel – *inimene ei saa / ei ole*. Erinevus seisneb vaid selles, et õppijakeeles on eelistatum modaalne predikaat, loomulikus keelekasutuses aga eksistentsiaalne.

Kolmas universaalne konstruktsioon on jaatava / eitava kõneliigi verbikeskne lausekonstruktsioon, milles on kasutatud sõna *inimene* mitmuse nominatiivi vormis: 1) ajakirjanduskeel – *need inimesed on*; 2) õppijakeel – *kõik / need inimesed on / hoidsid / tahavad*. Õppijakeeles on selles konstruktsioonis eelistatud täiendit *kõik*; samuti tuleb ette verbi leksikaalgrammatilist varieerumist (*on / hoidsid / tahavad*).

Neljas universaalne konstruktsioon on kvantorfraas, milles sõna *inimene* kasutatakse ainsuse partitiivi vormis: 1) ajakirjanduskeel – *ligi / veel / umbes 200 / kolm / 800 inimest*; 2) õppijakeel – *ainult / umbes kuus / sada inimest*. Nii loomulikus keelekasutuses kui ka õppijakeeles esineb kvantori leksikaalset varieerumist.

3.2. Erinevused

Erinevused loomuliku keelekasutuse ja õppijakeele vahel on seotud ajatähendusliku kvantorfraasiga, mis võib kuuluda

kaassõnafraasi. Selles konstruktsioonitüübis on EKI tekstikorpuse valimis eelistatud sõna *aeg* ainsuse nominatiivi, EVKKs aga ainsuse genitiivi: 1) ajakirjanduskeel – *mõni aeg hiljem ~ mõni aeg pärast <mida>*; 2) õppijakeel – *mõne aja pärast <tuli / oli>*. Tegu on ajatähendusliku kvantorfraasi kahe leksikaalgrammatilise variandiga, millest mõlemad on aktsepteeritavad ja vastavad eesti kirjakeele normile. Samas on loomulikus keelekasutuses eelistatum nominatiivi sisaldav konstruktsioon.

Teine erinevus ajakirjandus- ja õppijakeele vahel on seotud samuti ajatähendusliku kvantorfraasiga, milles sõna *aeg* kasutatakse ainsuse partitiivi vormis, kuid kvantorid varieeruvad: 1) ajakirjanduskeel – *väga pikka aega ~ juba tükk aega / kuu aega / nädal aega*; 2) õppijakeel – *väga vähe aega ~ nii / liiga palju aega*. Loomulikus keelekasutuses eelistatakse rõhusõnaga esile tuua pikemat ajaperioodi (*pikka aega, tükk aega, kuu / nädal aega*), samas kui õppijakeeles konstateeritakse kas ajanappust või seda, et aega on küllalt (*vähe aega* ja *palju aega*). Kvantorite varieerumine näitab, et ajatähendusliku kvantorfraasi kasutamine on õppijakeeles piiratum kui loomulikus keelekasutuses. Sedalaadi erinevuste väljatoomine on keeleõppe tarvis oluline, kuna need viitavad mõningatele vajakajäämistele õpikutes ja õppematerjalides. Viimaseid oleks vaja mitmekesistada, viies sisse uusi (ala)teemasid ja lisades õppeülesandeid, mida täites saab aktiveerida ka pikemat ajaperioodi väljendava ajatähendusliku kvantorfraasi kasutamist.

Kolmas erinevus on seotud konstruktsioonidega, milles sõna *inimene* on kasutatud mitmuse partitiivi vormis: 1) ajakirjanduskeel – eelistatud on dünaamilist lausekonstruktsiooni ja selle leksikaalgrammatilisi variante *inimesi kellel on ~ inimesi kes ei; neid inimesi kes*; *on inimesi kes*; 2) õppijakeel – eelistatud on rõhusõna sisaldavat kvantorfraasi *väga / nii palju inimesi*. Erinevad nii tekstiloomes aktiivselt kasutatavad konstrukt-

sioonitüübid kui ka esituslaad: kirjakeelele on omane jutustav-dünaamiline, õppijakeelt iseloomustab noomenisrtuktuuridele omane kirjeldav esituslaad. Niisugused erinevused näitavad kätte suuna, milles keeleõppija oskusi tuleks arendada.

Neljandat liiki erinevused, mis ilmnevad ajakirjandus- ja õppijakeeles eelistatud konstruktsioonide võrdlusest, viitavad sellele, et teatud konstruktsioonid varieeruvad õppijakeeles leksikaalgrammatiliselt tunduvalt rikkalikumalt kui loomulikus keelekasutuses. Ühelt poolt on näiteks eitava kõneliigi verbikeskne ainsuse nominatiivi sisaldav lausekonstruktsioon *inimene ei ole* / *ei saa* (vt eespool) loomuliku keelekasutuse ja õppijakeele universaal. Teisalt aga on õppijakeeles sama konstruktsiooni jaatava kõneliigi variandis esindatud rikkaliku leksikaalsemantilise varieeruvusega verbikeskne modaalne lausekonstruktsioon *inimene peab olema* / *maksma* / *teadma ~ inimene võib töötada*; *haritud (iga) inimene peab* / *saab* / *võib* / *tahab*, mis ajakirjanduskeele eelistuste seas pole statistiliselt esile tulnud. Ometi peaks eeldatavalt just hinnanguid sisaldav isiksuseline esituslaad olema omane publitsistlikule arutlusele ning modaalsed konstruktsioonid ajakirjanduskeelele. Samas pole ajakirjandustekstide analüüs seda kinnitanud. Tekib küsimus, miks? Kas korpusest tuleneva keeleanalüüsi andmeid saab interpreteerida kui meie päevalehtedele omaseks saanud tendentsi – vähe publitsistikat ja rohkesti isikupäratuid "lugusid"? Kas ajakirjanikud on hakanud hinnangute andmist vältima ning kirjutavad informatiivseid nupukesi? Samas on meeldiv tõdeda, et eesti keele õpetamisel on tehtud rõhuasetus oma arvamuse ja hinnangute väljendamisoskuse arendamisele – on ju õppijakeeles jaatava kõneliigi modaalne lausekonstruktsioon kõige sagedasem.

3.3. Kaks väljundit

EKI tekstikorpuse ja EVKK korpusest tulenev võrdlev uurimine suunab meie tähelepanu küsimustele, mis traditsioonilise korpuspõhise analüüsi puhul ei pruugi üles kerkida. Eelkõige puudutab see loomuliku keele kasutuseelistusi (konstruktsiooni tüüp, selle leksikaalgrammatiliste variantide olemasolu, rangelt valikuline käändevormide kasutus), mille taga võib aimata ja tänu millele saab selgemalt piiritleda mõningaid keelesüsteemi arengutendentse. Käesoleva uurimuse alusel sai kinnitust üldteada seisukoht, et loomuliku eesti keele tekstiloomes eelistatakse grammatilisi käändevorme ja valdavalt nominatiivi sisaldavaid konstruktsioone. Sama tendents ilmnes ka õppijakeele tekstiloomes. Tasub uurida, mis seda põhjustab. Kas keelesisene ökonoomia, suundumus käändeparadigma lühenemisele, mis lõppeks võib viia eesti keele nominatiivistumiseni? Või on tegu hoopis kontaktsituatsioonis tekkinud keelemuutusliku protsessiga, mida suunavad eesti keelt teise / võõrkeelena kõnelevad inimesed või eestlased, kes on omandanud teise ja kolmanda keele kõrgtasemel?

Teine tähelepanek seostub loomuliku eesti keele ja õppijakeele konstruktsioonierinevustega, millest annab teha järeldusi vajakajäämiste kohta keeleõppes. Nende kahe eesti keele variandi korpusest tulenev võrdlev analüüs võimaldab saada rangemapiirilise ettekujutuse sellest, mida õppijad on omandanud ning mida nad tegelikult peaksid omandama, et hakata eesti keelt kasutama samalaadselt eestlastega. Loomulikult jääb õhku küsimus, kas ikka kõik, mis ilmneb eestlase emakeelekasutuses, võib olla vaieldamatuks eeskujuks eesti keele kui teise / võõrkeele õppijale (nt jaatava kõneliigi modaalsete konstruktsioonide väljajäämine kasutuseelistuste hulgast). Lihtsaim moodus selles veendumiseks on eesti keele õigekirjaspelleri kasutamine ajakirjandustekstidel. Lisaks erinevuste väljatoomisele

tavapäraste konstruktsioonide kasutamisel loomulikus keelekasutuses ja õppijakeeles on siinkohal olulised loomuliku keelekasutuse stereotüüpsed konstruktsioonid, eriti hinnangu-, rõhumarkerite või sidususvahenditena.

Kirjandus

Baker, Paul 2006. Using corpora in discourse analysis. London: Continuum.

Behrens, Heike (Ed.) 2008. Corpora in language acquisition research: History, methods, perspectives. Amsterdam / Philadelphia: John Benjamins Publ. Co. Trends in language acquisition research 6.

Biber, Douglas 2004. Conversation text types: A multi-dimensional analysis. – JADT: 7es Journées internationales d´Analyse statistique des Données Textuelles / Ed. by G. Purnelle, C. Fairon, A. Dister. Louvain, 15–34. http://www.cavi.univ-paris3.fr/lexicometrica/jadt/jadt2004/pdf/JADT_000.pdf (9.05.2009).

Gabrielatos, Costas 2007. *If*-conditionals as modal colligations: A corpus-based investigation. – Proceedings of the Corpus Linguistics Conference: Corpus Linguistics 2007 / Ed. by M. Davies, P. Rayson, S. Hunston & Pernilla Danielsson. Birmingham: University of Birmingham. http://www.corpus.bham.ac.uk/corplingproceedings07/paper/256_Paper.pdf (24.03.2009).

Goldberg, Ada 1995. Constructions: a construction grammar approach to argument structure. Chicago: University of Chicago Press.

Eslon, Pille & Matsak, Erika 2009. Eesti keele kasutusvariandid: korpusest tulenev käändevormide võrdlev analüüs. – Eesti Rakenduslingvistika Ühingu aastaraamat 5 / Toim. H. Metslang, M. Langemets, M.-M. Sepper, R. Argus. Tallinn: Eesti Keele Sihtasutus, 79–110.

Eslon, Pille 2008. Käändevormide kasutussageduse võrdlus eesti õppijakeeles ja kirjakeeles. – Õppijakeele analüüs: võimalused, probleemid, vajadused / Toim. P. Eslon. Eesti filoloogia osakonna toimetised 10. Tallinn: Tallinna Ülikooli kirjastus, 31–66.

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

Granger, Sylviane 2003. Error-tagged learner corpora and CALL: A promising Synergy. – CALICO Journal 20 (3), 465–480.

Jantunen, Jarmo Harri 2009. Ei pelkästään mielikuvituksen puutteen vuoksi – Kieliaineistojen systemaattinen käyttö kielentutkimuksessa. – Virittäjä 1, 101–113.

Kilgi, Annika 2006. Mida eestlane teeb? – Oma Keel 12, 20–24.

Leech, Geoffrey 2004. Recent grammatical change in English: data, description, theory. – Advances in corpus linguistics: Papers from the 23rd international conference on English language research and computerized corpora (ICAME 23) Göteborg 22–26 may 2002 / Ed. by K. Aijmer & B. Altenberg. Amsterdam: Rodopi, 61–81.

Masini, Francesca 2005. Multi-word expressions between syntax and the Lexicon: the case of Italian verb-particle constructions. – SKY Journal of Linguistics 18, 145–173.

Muischnek, Kadri 2006. Verbi ja noomeni püsiühendid eesti keeles. Dissertationes philologiae Estonicae Universitatis Tartuensis 17. Tartu: Tartu Ülikooli Kirjastus.

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

Nikolajev, Alexander 2007. Suomen nominaalisen taivutusjärjestelman kvantitatiivista analyysia. – Kielitieteen päivät. Oulu 24.–25. toukokuuta 2007. Abstraktikirja, 75, http://www.oulu.fi/kielitieteenpaivat 2007/Abstraktit.pdf (25.03.2009).

Orpin, Debbie 2005. Corpus linguistics and critical discourse analysis: Examining the ideology of sleaze. – International Journal of Corpus Linguistics 10 / 1, 37–61.

Rayson, Paul 2002. Matrix: A statistical method and software tool for Linguistic analysis through corpus comparasion. Ph.D. thesis in Computer Science. Computing Department Lancaster University. http://ucrel.lancs.ac.uk/people/paul/publications/phd2003.pdf (30.04.2009).

Römer, Ute 2007. Learner language and the norms in native corpora and EFL teaching materials: A case study of English conditionals. – Anglistentag 2006. Halle. Proceedings / Ed. by S. Volk-Birke & J. Lippert. Trier: Wissenschaftlicher Verlag Trier, 355–363.

Stubbs, Michael 1996. Text and corpus analysis: Computer-assisted studies of language and culture. Language in Society 23. Oxford: Blackwell Publ.

Taylor, Charlotte 2008. What is corpus linguistics? What the data says. – ICAME Journal, 32: 179–200. http://icame.uib.no/ij32/ij32_179_200.pdf (24.03.2009).

Tognini Bonelli, Elena 2002. Functionally complete units of meaning across English and Italian: Towards a corpus-driven approach. – Lexis in Contrast. Corpus-based Approaches / Ed. by B. Altenberg & S. Granger. Philadelphia: John Benjamins, 73–95.

Contextual preferences the use of case forms in text production in Estonian

Pille Eslon

Summary

The objective of the study was to compare contextual preferences the use of case forms in two variants of Estonian – standard language (Standard Estonian corpus of the Institute of the Estonian Language) and learner language (Estonian interlanguage corpus of the Tallinn University). The material was most common nouns in Estonian *inimene* 'person', *sõna* 'word', *elu* 'life' ja *aeg* 'time'. The statistics showed frequency of constructions and grammatical case forms in Text. Statistical representation of lexical units, grammatical forms and constructions has enabled us to reveal such linguistic data

that, despite being characteristic of language use, is not easily accessible by traditional corpus-based analysis. Therefore we believe that corpus-driven studies using statistical and language software have a long-term theoretical and applied value (in language teaching and learning materials).

Analysis of contextual preferences making explicit the hidden tendencies that are working in the language system synchronically. For this purpose, the grammatical constructions and their lexico-grammatical variants (allowing only certain case forms) typical of Standard Estonian and learner Estonian were found out. Mere statistics of case forms would never have provided the data on their use in constructions and on their preference order in text production. Results of this study showed that in both language variants, the clearly preferred case is the nominative. Further research should reveal whether this phenomenon can be considered a process of nominativization, and what its possible motivations could be.

Keywords: corpus linguistics, corpus-driven comparative language analysis, use of case forms in most frequently grammatical constructions, Estonian language

Autor

PhD Pille Eslon, Tallinna Ülikooli eesti keele ja kultuuri instituudi vanemteadur, riikliku programmi „Eesti keele keeletehnoloogia tugi (2006–2010)" projekti „VAKO – Eesti vahekeele korpuse keeletarkvara ja keeletehnoloogilise ressursi arendamine" vastutav täitja, pille.eslon@tlu.ee

SYNTAKTISESTI KOODATTU OPPIJANKIELEN KORPUS: MAHDOLLISUUKSIA JA KYSYMYKSIÄ

Ilmari Ivaska, Kirsti Siitonen

Abstrakti

Oppijankieleen kohdistuva korpuslingvistiikka on kasvava tutkimusala, joka voi kertoa oppijankielestä paljon sellaisia asioita, jotka ovat aiemmin olleet tutkijoiden ulottumattomissa. Syntaktisesti koodattu korpus laajentaa näitä tutkimusmahdollisuuksia entisestään. Syntaktisesti koodatun korpuksen kehittäminen on kuitenkin pitkä prosessi, jonka laatiminen herättää runsaasti kysymyksiä kaikissa sen työstämisen vaiheissa. Tässä artikkelissa esittelemme Turun yliopiston Lauseopin arkiston osaksi tulevaa Edistyneiden suomenoppijoiden korpuksen (LAS2) toteutusta sekä sen kehittämisessä ilmi tulleita kysymyksiä ja valittuja ratkaisuja. Korpuksen ensimmäinen osa on hakusanoitettu ja se on koodattu niin morfologisesti kuin syntaktisestikin, minkä lisäksi korpukseen liitetään virhekoodaus. Korpus mahdollistaa myös informanttikohtaisen pitkittäistutkimuksen.

Avainsanat: kielen koodaus, korpuslingvistiikka, suomi toisena kielenä, syntaksi

1. Johdanto

Turun yliopiston edistyneiden suomenoppijoiden kielestä koostuvan korpuksen (LAS2) koodaus on aloitettu syksyllä 2008. Koodaus toteutetaan Turun yliopiston Lauseopin arkiston mallin mukaisesti ja aineisto muokataan XML-muotoon (”Extensible Markup Language”, ks. alempana). Korpus tullaan liittämään osaksi Turun yliopiston Lauseopin arkistoa, jossa se on tutkijoiden käytettävissä verkkoselaimen välityksellä.

Koodaaminen tehdään vaiheittain kumuloituvasti ja siitä pyritään tekemään mahdollisimman joustava niin, että edeltävä vaihe luo pohjaa seuraavalle. Ensimmäisessä vaiheessa tarkoituksena on keskittyä morfologisen tason ja sanatason koodaukseen merkitsemällä aineistoon runsaasti morfologista informaatiota kustakin sanasta. Toisessa vaiheessa morfologinen koodaus kontekstualisoidaan, aineisto rakenteistetaan ja siihen merkitään syntaktista informaatiota. Rinnan koodauksen eri vaiheiden kanssa merkitään kaikki standardoidusta yleiskielestä poikkeavat muodot ja käyttöyhteydet kommenttikoodilla. Turun yliopiston Edistyneiden suomenoppijoiden korpus seuraa pääpiirteissään Nobufumi Inaban kehittämää syntaktisesti koodatun korpuksen tekotapaa (Ks. Inaba 2007: 147–161).

2. Koodatun oppijakorpuksen laatiminen

2.1. Aineisto ja tekstilaji

Aineistona on Turun yliopiston suomen ja sen sukukielten maisteriohjelman opiskelijoiden kirjoittamaa opiskeluun liittyvää asiaproosaa. Seuranta-aika on 2–3 vuotta. Opiskelijoiden suomen kielen taito on jo alussa vahva. Lisäksi korpuksen osaksi kootaan ensikielisten suomen kielen opiskelijoiden tenttivastauksista koostuva vertailuaineisto.

Tutkittavana on monipuolinen tekstilajivalikoima, joka sisältää tenttivastauksia, esseitä ja tutkielmia sekä katsauksia ja raportteja. Korpuksen laatimisen ensimmäisessä vaiheessa pääpaino on tenttivastauksissa. Aineisto karttuu jatkuvasti. Nykyisellä aineistonkeräämismetodilla korpuksen tenttivastausten ja esseiden vuotuinen kertymä on 15 000–20 000 saneen luokkaa kummassakin osiossa, tutkielmien osalta se on noin 60 000 sanetta. Katsauksia ja raportteja tulee vain muutama vuosittain, noin 10 000 saneen luokkaa.

Tekstilajilla on luonnollisesti suuri merkitys aineiston laatuun. Esimerkiksi tenttivastauksista voidaan olettaa, että niiden muoto on osittain tenttikysymyksen muodon inspiroima. Ne lienevät myös jossakin määrin stereotyyppisiä niin, että esimerkiksi eri vastausten alut voivat olla hyvinkin samankaltaisia. Kuten Jyrki Kalliokoski huomauttaa, tekstilaji on ennen kaikkea sosiokulttuurinen käsite (Kalliokoski 2006: 240). Tekstilajin hallinta on siis niin ikään kielen omaksumiseen kuuluvaa stereotyyppistä tietoa siitä, miten kielenkäyttäjän on tapana toimia kussakin tilanteessa. Se on osa kielellistä sujuvuutta, jossa keskeistä on variaatio (Mt., 248). Sylvaine Grangerin mukaan korpusten avulla onkin mahdollista tarkastella kielen variaation ja eri tekstilajien yhteyksiä (Granger 2002: 4–5). Vertailuaineiston avulla tällaisten tekstilajikohtaisten ilmiöiden tarkasteleminen on mahdollista.

2.2. Morfologinen koodaus ja sanakirja

Aineisto muokataan, kuten jo johdannossa todettiinkin, XML-muotoon. Teknisesti tämä tarkoittaa html-kielen kaltaisten tagien käyttöä, joiden avulla aineisto rakenteistetaan hierarkkisesti. Koodaus on Kotimaisten kielten tutkimuskeskuksessa Kotuksessa kehitetyn TEI-ohjeistuksen (*The Text Encoding Initiative*) muunnoksen mukainen (Inaba 2007: 151).

Morfologisessa koodauksessa aineistoon merkitään seuraavat tiedot: oikeakielinen hakusanamuoto silloin, kun se on mahdollista (<lemma>[1]); sanaluokka (<pos>); muoto-opilliset metatiedot silloin, kun niiden tulkitseminen on mahdollista (<mrp>); virhekommentit silloin, kun niiden tunnistaminen on mahdollista (<com>). Morfologisista metatiedoista koodataan sanaluokasta riippuen seuraavat seikat: sijamuoto, komparaatiomuoto, luku, omistusliitteet, pääluokka, tempus, modus, persoona ja finiittisyys. Virhekommentoinnissa voidaan tässä vaiheessa huomioida seuraavat ilmiöt: vartalovirheet, astevaihteluvirheet, morfologiset taivutusvirheet ja sanojen sekoittuminen. Esimerkiksi virheellisesti muodostetun *hyvä*-adjektiivin vertailumuoto *hyvempi* koodataan sanakirjavaiheessa seuraavasti:

<w lemma=”hyvä” pos=”a” cmp sg nom" com="virheellinen cmp">hyvempi</w>[2]

Tätä työskentelyvaihetta kutsutaan sanakirjan tekemiseksi. Sanakirjaa tehtäessä koko aineistoa käsitellään sanamuodoittain aakkosjärjestyksessä. Kukin sanamuoto koodataan vain kerran, mistä syystä homonymiatapaukset koodataan kaikin mahdollisin tulkinnoin. Tällöin todennäköisempi tulkinta koodataan sanan morfologiseksi koodaukseksi ja epätodennäköisempi sanan kommentiksi. Tarpeeton koodaus poistetaan sitten, kun koodaus kontekstualisoidaan. Kun esimerkiksi *tulla*-verbin konnegaatiomuoto ja imperatiivimuoto *tule* lankeavat yhteen, on sanan koodaus sanakirjavaiheessa seuraavanlainen:

---

[1] Sulkeissa oleva merkintä kertoo kunkin seikan kohdalla korpuksessa käytettävän XML-tagin muodon.

[2] w = word, hyvä = hakusana, a = adjektiivi, cmp = komparaatio, sg = yksikkö, nom = nominatiivi, virheellinen cmp = virheellisesti muodostettu komparaatio.

<w lemma="tulla" pos="v" mrp="conneg ind pres" com="fin impv pres sg2">tule</w>[3]

Sanakirjan tekeminen on osa koodauksen automaattistamista. Kun kukin sanamuoto on koodattu kerran, samaa morfologista koodausta voidaan käyttää korpuksen karttuessa myös jatkossa. Kun korpusta kasvatetaan lisäämällä siihen uutta digitalisoitua aineistoa, ajetaan uusi aineisto ensimmäisessä vaiheessa sanakirjan kautta. Tämän jälkeen sanakirjaan koodataan ainoastaan ne sanamuodot, joita se ei ennestään sisältänyt. Korpuksen kasvaessa ja sanamuotojen lisääntyessä morfologinen raakakoodaus hioutuu siis lähes automaattiseksi.

2.3. Syntaktinen koodaus

Syntaktisessa koodauksessa morfologinen koodaus kontekstualisoidaan ja aineisto rakenteistetaan informanteittain tekstintuottamisajankohdan, tekstikokonaisuuksien, kappaleiden, virkkeiden ja lauseiden osalta. Samalla aineistoon lisätään syntaktisten funktioiden koodaus ja virhekommentointiin lisätään syntaktisia seikkoja koskevat huomautukset.

Syntaktisen koodauksen aluksi valmis sanakirja syötetään takaisin aineistoon eli morfologinen koodaus kontekstualisoidaan. Koska aineisto on digitalisoitu virkkeittäin ja siihen on merkitty kappaleet ja tekstikokonaisuuksien alkukohdat, voidaan virkkeet (<s>), kappaleet (<p>), tekstikokonaisuudet (<teksti>) sekä tekstintuottoajankohdat (<tentti> ja <paivamaara>) rakenteistaa automaattisesti skriptien eli komentosarjojen avulla. Tämä koodaus ei ole virheetön, mutta se no-

[3] tulla = hakusana, v = verbi, conneg = konnegaatiomuoto, ind = indikatiivi, pres = preesens, fin = finiittimuoto, impv = imperatiivi, sg2 = yksikön 2. persoona.

peuttaa työskentelyä, sillä virheet voidaan korjata käsin kontekstisidonnaista koodausta tehtäessä.

Kontekstisidonnainen koodaus tarkoittaa aineiston jakamista lauseisiin (<cl>) ja sanojen syntaktisen roolin (<fun>) koodausta sekä lauseiden tyypittelemistä myönteisiin ja kielteisiin väite- ja kysymyslauseisiin (<cl type>). Samalla aineistosta poistetaan sanakirjan aiemmin mainitut päällekkäiset morfologiset koodaukset sekä muut vastaan tulevat koodausvirheet. Asetelmissa 1 ja 2 on virke *Merkitys on myös sellainen, että sana ei voi olla kovin vanha, lisäksi ovat vastineet indoeurooppalaisissa kielissä.* ennen kontekstisidonnaista koodausta ja sen jälkeen.

Asetelma 1. Korpuksen virke *Merkitys on myös sellainen, että sana ei voi olla kovin vanha, lisäksi ovat vastineet indoeurooppalaisissa kielissä.* ennen kontekstuaalista koodausta.

```
<s num="9">
<cl type="" fun="" com="">
<w lemma="merkitys" pos="n" mrp="sg nom" fun="" com="">Merkitys</w>
<w lemma="olla" pos="v" mrp="fin ind pres sg3" fun="" com="">on</w>
<w lemma="myös" pos="adv" mrp="" fun="" com="">myös</w>
<w lemma="sellainen" pos="a" mrp="sg nom" fun="" com="">sellainen</w>
<w lemma="" pos="" mrp="" fun="" com="">,</w>
<w lemma="että" pos="cnj" mrp="" fun="" com="">että</w>
<w lemma="sana" pos="n" mrp="sg nom" fun="" com="">sana</w>
```

```
<w lemma="ei" pos="neg" mrp="sg3" fun="" com="">ei</w>

<w lemma="voida" pos="v" mrp="conneg ind pres " fun=""
com="fin ind pres sg3">voi</w>

<w lemma="olla" pos="v" mrp="inf1" fun=""
com="">olla</w>

<w lemma="kovin" pos="adv" mrp="" fun=""
com="">kovin</w>

<w lemma="vanha" pos="a" mrp="sg nom" fun=""
com="">vanha</w>

<w lemma="" pos="" mrp="" fun="" com="">,</w>

<w lemma="lisäksi" pos="p:post" mrp="" fun=""
com="">lisäksi</w>

<w lemma="olla" pos="v" mrp="fin ind pres pl3" fun=""
com="">ovat</w>

<w lemma="vastine" pos="n" mrp="pl nom" fun=""
com="">vastineet</w>

<w lemma="indoeurooppalainen" pos="a" mrp="pl ine"
fun="" com="">indoeurooppalaisissa</w>

<w lemma="kieli" pos="n" mrp="pl ine" fun=""
com="">kielissä</w>

<w lemma="" pos="" mrp="" fun="" com="">.</w>

</cl>

</s>
```

Virke on kontekstuaalisesti koodattaessa jaettu kolmeen lauseeseen ja kunkin lauseen alkuun on merkitty lauseen tyyppi. Kuhunkin sanaan on koodattu sen syntaktinen funktio, toisen

lauseen konnegaatiomuotoisesta verbistä *voi* on poistettu vaihtoehtoinen koodaus ja kolmannen lauseen predikaattiin ja subjektiin on lisätty kommentti niiden odotuksenvastaisesta muodosta. (Ks. asetelma 2).

Asetelma 2. Korpuksen virke *Merkitys on myös sellainen, että sana ei voi olla kovin vanha, lisäksi ovat vastineet indoeurooppalaisissa kielissä.* kontekstuaalisen koodaamisen jälkeen. Koodauksessa tehdyt muutokset on lihavoitu.

<s num="9">

<cl type="**affdecl**" fun="" com="">

<w lemma="merkitys" pos="n" mrp="sg nom" fun="**npsubj**" com="">Merkitys</w>

<w lemma="olla" pos="v" mrp="fin ind pres sg3" fun="**pred**" com="">on</w>

<w lemma="myös" pos="adv" mrp="" fun="**advl**" com="">myös</w>

<w lemma="sellainen" pos="a" mrp="sg nom" fun="**compl:s**" com="">sellainen</w>

<w lemma="" pos="" mrp="" fun="" com="">,</w>

</cl>

<cl type="negdecl" fun="" com="">

<w lemma="että" pos="cnj" mrp="" fun="**lauseyhd**" com="">että</w>

<w lemma="sana" pos="n" mrp="sg nom" fun="**npsubj**" com="">sana</w>

<w lemma="ei" pos="neg" mrp="sg3" fun="**pred**" com="">ei</w>

<w lemma="voida" pos="v" mrp="conneg ind pres" fun="**pred2**" com=**POISTO[""]**>voi</w>

<w lemma="olla" pos="v" mrp="inf1" fun="**pred3**" com="">olla</w>

<w lemma="kovin" pos="adv" mrp="" fun="**nmod**" com="">kovin</w>

<w lemma="vanha" pos="a" mrp="sg nom" fun="**compl:s**" com="">vanha</w>

<w lemma="" pos="" mrp="" fun="" com="">,</w>

</cl>

<cl type="affdecl" fun="" com="">

<w lemma="lisäksi" pos="p:post" mrp="" fun="**advl**" com="">lisäksi</w>

<w lemma="olla" pos="v" mrp="fin ind pres pl3" fun="**pred**" com="**kongr_e**">ovat</w>

<w lemma="vastine" pos="n" mrp="pl nom" fun="**npsubj**" com="**sija_e_subj**">vastineet</w>

<w lemma="indoeurooppalainen" pos="a" mrp="pl ine" fun="**nmod**" com="">indoeurooppalaisissa</w>

<w lemma="kieli" pos="n" mrp="pl ine" fun="**advl**" com="">kielissä</w>

<w lemma="" pos="" mrp="" fun="" com="">.</w>

</cl>

</s>

Kuten aiemmin mainittiin, kieliopillisen informaation lisäksi aineisto identifioidaan informanttikohtaisesti tekstintuottamisajankohdan perusteella. Tämä metodi mahdollistaa pitkittäistarkastelun tekemisen.

Syntaktinen koodaus on huomattavasti vaativampaa kuin morfologinen koodaus. Lauseoppia käsittelevässä tutkimuksessa esitetään hyvin paljon keskenään toisistaan poikkeavia näkemyksiä. Syntaktisessa koodauksessa onkin tehtävä paljon enemmän itsenäisiä päätöksiä ja valintoja kuin morfologisessa koodauksessa. Näin ollen korpuksen syntaktista koodausta aloitettaessa eri ratkaisuista on keskusteltava tarkoin ja työn edetessä kaikki ratkaisut on huolellisesti kirjattava. Tärkein tavoite on korpuksen yhdenmukaisuus ja tehtyjen ratkaisujen transparenttius. Esimerkiksi *haluan tehdä* -tyyppiset verbiketjut on syntaktiselta funktioltaan koodattu Ison suomen kieliopin tulkinnan mukaisesti verbiketjuiksi (koodeina pred, pred2 jne) (ISK: 493–495). Samoin kieltomuodot, kuten *ei voi olla*, koodataan syntaktisesti verbiketjuiksi (pred, pred2, pred3). Muita tulkinnanvaraisia lauseopillisia ilmiöitä ovat esimerkiksi: *meistä tulee opettajia* kaltaisen tuloslauseen subjektipaikkaisen jäsenen *opettajia* syntaktinen funktio, mikä on korpuksessa koodattu subjektinpredikatiiviksi (compl:s); lauseensisäisen infinitiivilausekkeen jäsentäminen, missä on päädytty tulkitsemaan lausekkeen pääsanan edustavan kulloinkin kyseessä olevaa lauseenjäsentä ja muiden jäsentyvän osana infinitiivilauseketta, esim. *Sen* (nmod) *tehtävänä* (advl:p) *on* (pred) *analysoida* (infsubj) *virheitä* (npobj)[4].

Korpukseen tullaan jatkossa soveltamaan Mikael Agricolan teosten tieteellinen editio ja morfosyntaktinen tietokanta -tut-

[4] nmod = substantiivin määrite, advl:p = predikatiiviadverbiaali, pred = predikaatti, infsubj = infinitiivisubjekti, npobj = substantiiviobjekti.

kimushankkeessa kehitettyä koodausjärjestelmää. Inaban mukaan syntaksin puoliautomaattinen koodaus on näillä metodeilla sanaluokasta riippuen parhaimmillaan 80–90 prosentin luokkaa. Adverbien ja adpositioiden osalta koodauksen tarkkuus on yli 90 prosenttia ja adjektiivien osalta noin 80 prosenttia. Substantiivien kohdalla tarkkuus laskee alle 60 prosentin (Inaba 2009). Osittaisesta automatisoimisesta huolimatta koko aineisto tullaan jatkossakin käymään läpi myös manuaalisesti koodauksen tarkistamiseksi ja rakenteistamiseksi, virheiden korjaamiseksi ja ylimääräisten koodirivien karsimiseksi (ks. esim. Inaba 2007). Automatisoitavien prosessien avulla koodaustyö kuitenkin kevenee huomattavasti, mikä mahdollistanee jatkossa koodatun aineiston nopeamman kasvattamisen ja korpuksen luotettavuuden parantamisen.

2.4. Virhetyypittely

Korpukseen merkitään myös virhekoodaus. Tämä koodaus ja siihen liittyvä virhetyypittely tehdään aiemmin kuvattujen koodausvaiheiden aikana tehdyn virhekommentoinnin pohjalta. Näin varmistetaan seuraavassa kappaleessa tarkemmin esiteltävä aineiston laadun kannalta relevantti luokittelu.

Morfologisen ja kontekstuaalisen koodauksen jälkeen niissä merkityt kommentit aineiston normipoikkeamista ja kohdekielen kannalta epätyypillisistä kielellisistä ratkaisuista kerätään yhteen. Näiden kommenttien pohjalta pystytään hahmottamaan aineiston kannalta tarkoituksenmukaisia virheryhmiä niin niiden laadun kuin esiintymien määränkin puolesta. Tämän tyypittelyn perusteella Turun yliopiston Lauseopin arkiston Edistyneiden suomenoppijoiden korpukseen koodataan viisiportainen hierarkkinen virhetyyppiluokitus.

Luokituksen ensimmäinen taso on kunkin tekijän norminmukaisuus tai normipoikkeama. Tämän jälkeen tulee tyypittelyn kuusi pääluokkaa, jotka ovat seuraavat:

1) sanastolliset virheet
2) sanastollis-johto-opilliset virheet
3) sanastollis-morfologiset virheet
4) morfologiset virheet
5) syntaktiset virheet
6) lauserakenteelliset virheet.

Luokittelu on väistämättä joiltakin osin ristikkäinen, sillä monet aineiston normipoikkeamat ovat joiltakin osiltaan monitasoisia. Korpuksessa käytettävässä tyypittelyssä on keskeistä se, että luokitus on toimiva ainoastaan yhdessä korpuksen muun koodauksen kanssa. Näin esimerkiksi objektin sijavalinnassa esiintyvä normipoikkeama merkitään virhekoodauksessa ainoastaan sijavirheenä, mutta koska sanaan on koodattu sen syntaktinen funktio, korpuksesta pystytään tarpeen mukaan tarkastelemaan niin yleistä objektien sijavariaatiota, objektien herkkyyttä normipoikkeamille kuin objektien sijassa esiintyviä normipoikkeamiakin. Virhetyypittelyä ja sen teknistä toteutusta ollaan parhaillaan kehittämässä.

2.5. Subjektiivisen tulkinnan ongelma

Oppijankieltä käsittelevää, syntaktisesti koodattavaa korpusta tehtäessä kohdataan väistämättä lukuisia koodaajien subjektiiviseen tulkintaan liittyviä ongelmia, joista osaa on sivuttu tässä artikkelissa jo aiemmin. Kuten sanojen syntaktisten funktioiden koodaamista käsiteltäessä todettiin, tärkeintä korpuksen käytettävyyden kannalta on tehtyjen ratkaisujen yhdenmukaisuus ja niiden tarkka ja transparentti dokumentointi ja perustelu.

Suurin ja monitulkintaisin ongelma korpuksessa on aiemmin esitelty virhetyypittely ja sen liittäminen osaksi koodausta. Virheen käsite on hyvin monisyinen. Hankkeessa työskentelevät ovat erittäin tietoisia siitä, että välikielen poikkeamia ei pidetä virheinä ja että sellaisessakin kielitaitonäkemyksessä, jossa jokin ilmiö luokitellaan virheeksi, ilmiö ei välttämättä ole virhe kaikissa rekistereissä. Korpuksen myöhempää käyttöä varten on kuitenkin pidetty järkevänä, että äidinkielisen koodaajan kieli-intuition vastaiset ilmaukset merkitään. Myöhemmin korpusta tutkimuksessaan käyttävä ratkaisee suhteensa merkintöihin tapauskohtaisesti. Ennen korpuksen virhekoodauksen julkistamista on välttämätöntä, että sen tarkistaa koodaajan lisäksi vähintään yksi ensikielenään suomea puhuva hankkeen työntekijä.

3. Korpuksen soveltaminen suomi toisena kielenä -tutkimukseen

Syntaktisesti koodattu oppijankorpus pystyy tarjoamaan uutta tietoa sellaisista ilmiöistä, joita on aiemmin ollut työlästä tai vaikeaa tutkia kvantitatiivisin metodein. Syntaktinen koodaus mahdollistaa esimerkiksi lauserakenteiden esiintymisfrekvenssien tarkastelun suhteessa ensikielisten suomenpuhujien lauserakenteisiin sekä tällaisten ilmiöiden esiintymisehtojen ja niiden valintapreferenssien tarkastelun niin itse oppija-aineistossa kuin korpuksen osaksi liitettävässä vertailuaineistossakin (vrt. esim. Jantunen 2004: 15, 29). Keskeistä syntaktisessa koodauksessa on lisäksi se, että sen avulla pystytään luontevasti nostamaan tarkastelun kohteeksi nimenomaan lauseet pelkkien muotojen asemesta. Kuten Granger kirjoittaa, korpusten teho on nimenomaan kielen frekvenssejä tarkasteltaessa, sillä se on ala, josta kielenpuhujilla on hyvin vähän intuitiivista tietoa (Granger 2002: 4).

Hakusanoitettu eli lemmattu aineisto mahdollistaa periaatteessa myös monipuolisen sanastontutkimuksen. Toistaiseksi korpuksen koodatun aineiston laatu kuitenkin rajoittaa tätä tutkimusta, sillä kielitieteeseen painottuvat tenttivastaukset eivät anna todenmukaista kuvaa kielenoppijoiden sanastosta. Kun korpusta laajennetaan myös muihin tekstilajeihin, tämä on mahdollista.

Jarmo Jantunen kirjoittaa sellaisista oppijankielen universaaleista, joiden tutkimiseen korpustutkimus tarjoaa uusia keinoja ja näkökulmia. Näitä ovat hänen mukaansa kielenainesten epätyypilliset frekvenssit, kieltenvälinen vaikutus, yksinkertaisuus ja epäkonventionaalisuus (Jantunen 2008: 70–81). Voidaankin laajasti ottaen ajatella, että syntaktisesti koodattu oppijankielenkorpus mahdollistaa näiden seikkojen tutkimuksen myös lauseopillisten seikkojen osalta. Suurista aineistoista on mahdollista myös nähdä, minkälaisia ensikielen puhujille tyypillisiä piirteitä tai ilmiöitä ei ollenkaan esiinny oppijoiden teksteissä. Näin mahdollistuu siis myös hyvin hankalasti havaittavan välttämisen (*avoidance*) tutkiminen.

Edellä esitetyn lisäksi voidaan korpuksen teksteistä tutkia ylipäänsä kaikkea sellaista, mitä muustakin S2-aineistosta voidaan tutkia. Tutkimusta vain helpottaa se, että kaikki materiaali on valmiiksi järjestettyä ja digitaalisessa muodossa.

Tärkeä seikka korpusta käytettäessä on sen tulevan virhekoodauksen soveltaminen. Virhekoodausta on syytä tulkita kriittisesti eikä sitä voi missään tapauksessa käyttää tutkimushypoteesien suorana toteennäyttäjänä. Virhekoodauksen perusteella ei esimerkiksi ole syytä laskea suoraan virhefrekvenssejä. On muistettava, että virhekoodaus on ennen kaikkea tutkimuksen lähtökohta, ei sen tulos. Sen avulla pystytään osoittamaan laajasta aineistosta edistyneiden oppijoiden suomelle

ominaisia piirteitä, jotka tuntuvat ensikielisen suomenpuhujan intuition vastaisilta. Tämä on kuitenkin vasta lähtökohta, jolle tutkimus voidaan perustaa.

4. Lopuksi

Koodauksen tässä vaiheessa haasteena on hahmottaa kokonaisuus kussakin vaiheessa siten, että tehty työ ei rajoita tulevia, toistaiseksi osin huomaamattomia oivalluksia. Parhaimmassa tapauksessa nyt tehtävä työ helpottaa tulevaa ilman, että siitä koituu varsinaista lisävaivaa missään vaiheessa. Aineisto koodataan lauseopillisesti ja sitä kasvatetaan jatkuvasti niin määrällisesti kuin laadullisestikin. Monipuolinen pitkittäisaineisto tarjoaa ajallisen perspektiivin kunkin informantin tuotoksiin. Sen lisäksi arkisto sisältää lukuisia eri tekstilajeja samoilta informanteilta. Kun nämä seikat saadaan mukaan ja näkyviin korpuksen kokonaisuuteen, korpuksen käytettävyys ja monipuolisuus paranee entisestään.

Arkistoon pyritään tallentamaan myös saman informantin saman tekstin eri versioita, joiden avulla voitaneen tavoittaa esimerkiksi pedagogisesti kiinnostavaa tietoa korjaamisesta sekä fossilisoituneista rakenteista.

Kirjallisuus

Granger, Sylviane 2002. A Bird's-eye view of learner corpus research. – Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching / Toim. S. Granger & J. Hung & S. Petch-Tyson. Philadelphia: John Benjamins, 3–37.

Inaba, Nobufumi 2007. Mikael Agricolan teokset tietokannan muodossa. – Agricolan aika / Toim. K. Häkkinen & T. Vaittinen. Helsinki: BTJ, 147–161.

Inaba, Nobufumi 2009. Re: Koodaamisen automatisoiminen. [Sähköpostiviesti 11.3.2009]

ISK = Hakulinen, Auli & Vilkuna, Maria & Korhonen, Riitta & Koivisto, Vesa & Heinonen, Tarja Riitta & Alho, Irja 2004. Iso suomen kielioppi. SKST 950. Helsinki: Suomalaisen Kirjallisuuden Seura.

Jantunen, Jarmo 2004. Synonymia ja käännössuomi: korpusnäkökulma samamerkityksisyyden kontekstuaalisuuteen ja käännöskielen leksikaalisiin erityispiirteisiin. Joensuu: Joensuun yliopiston humanistisia julkaisuja.

Jantunen, Jarmo 2008. Haasteita oppijankielen analyysille: oppijankielen universaalit. – Õppijakeele analüüs: Võimalused, probleemid, vajadused / Toim. P. Eslon. Eesti filoloogia osakonna toimetised 10. Tallinn: TLÜ Kirjastus, 67–91.

Kalliokoski, Jyrki 2006. Tekstilajin taju ja toisella kielellä kirjoittaminen. – Genre – tekstilaji / Toim. A. Mäntynen & S. Shore & A. Solin. Tietolipas 213. Helsinki: Suomalaisen Kirjallisuuden Seura, 240–265.

Syntactically encoded corpus of learner language: Opportunities and challenges

Ilmari Ivaska, Kirsti Siitonen

Summary

The encoding of the corpus of advanced Finnish began in autumn 2008 at the University of Turku. The corpus is being modified into the XML-format (”Extensible Markup Language) and it follows the TEI-directions (The Text Encoding Initiative) given by the Research Institute for the Languages of Finland.

The encoding is done cumulatively in order to retain the flexibility of the corpus throughout the progress of the project. The first stage is to focus on the morphological and lexical level. All word types are encoded alphabetically with various morphological information. The lemmatisation of the material is executed synchronously. All the word types are encoded only once and the encoding is, then, extended to every token. This method is called the compiling of the corpus' dictionary.

The second stage is the syntactical encoding, in which the morphological encoding is contextualised separately for each informant. In this phase, the corpus is organised by the date of the texts, text entities, paragraphs, sentences and clauses. Simultaneously, each token is encoded by its syntactical function.

Alongside these steps, the material is also commented in the respect of errors and unidiomatic formal and lexical solutions. In the third step the comments are collected and sorted to generate a corpus-based error tagging system. In these manners the corpus will be labelled with a hierarchical error tagging. However, the concept of an error is highly complex and, thus, the error tagging should be considered solely as the basis of every particular study.

Syntactically encoded learner language corpus can provide new knowledge about the phenomena that have so far remained too demanding to reach with non-computer based quantitative methods. It enables for example the studies of the frequencies of syntactical structures in advanced learner's Finnish.

Keywords: encoding, corpus linguistics, Finnish as a second language, syntax

Esittely

HuK Ilmari Ivaska on Turun yliopiston edistyneiden suomenoppijoiden korpus -hankkeessa LAS2 työskennellyt alusta (kevät 2007) alkaen ja on kirjoittamassa parhaillaan pro gradu -työtään korpuksen aineistosta, itivas@utu.fi

FT Kirsti Siitonen, ma prof. on opettanut vuodesta 1976 suomea vieraana tai toisena kielenä eri yliopistoissa Suomessa ja ulkomailla. Hän on erityisesti perehtynyt edistyneiden suomenoppijoiden kieleen ja julkaissut alalta vuonna 1999 väitöskirjansa. Kun Turun yliopistoon perustettiin vuonna 2005 suomen ja sen sukukielten maisteriohjelma, tarjoutui tilaisuus kerätä jatkuvasti kertyvää edistyneiden suomenoppijoiden kirjoittamaa tekstiä. Siitonen perusti v. 2007 korpushankkeen, jonka johtaja hän on, kisiito@utu.fi

ÕPPIJAKEELE KORPUSANALÜÜSI TÄIENDAVATEST MEETODITEST

Annekatrin Kaivapalu

Ülevaade

Õppijakeele korpusanalüüsi keskmes on esmajoones õppija keeleline produktsioon, mida on valdavalt uuritud veaanalüüsi meetodi abil. Seevastu sihtkeele produtseerimisprotsessi kohta saadud teave on juhuslikum või puudub üldse. Artiklis keskendutakse veaanalüüsi peamistele probleemidele nagu vigade klassifitseerimine ja interpreteerimine ning vigade psühholingvistiline reaalsus. Samuti kirjeldatakse meetodeid, mis õppijakeele korpusanalüüsi selles osas täiendavad. *Eesti vahekeele korpuse* sõnajärjevigade näitel analüüsitakse tajutesti kasutamisvõimalusi vigade psühholingvistilise reaalsuse väljaselgitamisel nii emakeelse keelekasutaja kui ka keeleõppija seisukohalt. Lisaks kirjeldatakse õppijakeele korpusanalüüsi täiendavaid introspektiivseid meetodeid (valjusti mõtlemine ja intervjuu) ning tutvustatakse kirjutamisprotsessi fikseerivat arvutiprogrammi ScriptLog[1].

Võtmesõnad: korpuspõhine veaanalüüs, vigade psühholingvistiline reaalsus, eesti keele sõnajärjemallid, eestikeelne keele-

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

kasutaja, venekeelne eesti keele õppija, introspektiivsed meetodid

1. Õppijakeele korpusanalüüs

Õppijakeele uurimine mahukate elektrooniliste andmekogude, õppijakeele korpuste[2] abil on muutumas üha levinumaks meetodiks keeleõppealases uurimistöös. Õppijakeele korpusi koostatakse ja arendatakse mitmel pool maailmas[3] ning järjest lisandub projekte, mille eesmärk on uurida korpusanalüüsi abil õppijakeele keeleomaseid ja universaalseid jooni nagu keeltevaheline mõju (Kaivapalu 2008; Tenfjord 2008), keelendite ebatüüpiline sagedus, lihtsustamine, ebatüüpiline konventsionaalsus jt (Jantunen 2007). Ka keeleoskuse omandamise ja mõõtmise uurimist ühendav mitteformaalne SLATE-võrgustik[4] kasutab ainestikuna esmajoones õppijakeele korpusi, nii õppetöö käigus loodud kirjalike tekstide kui ka keeleoskuse tasemeid mõõtvate testide elektroonilisi kogusid (Tarnanen 2007).

Korpuspõhise lähenemisviisi eeliseks võrreldes traditsioonilise uurimisega on kahtlemata tulemuste parem üldistatavus: suurte andmehulkade analüüs annab objektiivsema pildi eri tasemel õppijate keeleoskuse arengust ning keele produtseerimise spetsiifikast. Samas on korpuspõhisel lähenemisviisil mitmeid piiranguid: korpusainestik ja -metodoloogia ei võimalda

---

[2] S. Grangeri (2003: 465) määratluse kohaselt on paralleelselt õppijakeelekorpuse mõistega kasutusel terminid *vahekeelekorpus* (ingl *interlanguage corpus*) ja *teise keele korpus* (ingl *L2 corpus*).

[3] Ülevaadet õppijakeele korpustest vt Eslon & Metslang 2007, teoksil olevatest uurimisprojektidest vt http://cecl.fltr.ucl.ac.be/SummerSchool2007/participants%20research%20interests.html (19.8.2009).

[4] http://www.jyu.fi/hum/laitokset/kielet/cefling/suom (19.8.2009).

uurida sugugi kõiki õppijakeele aspekte. Olulisim on asjaolu, et õppijakeele korpusanalüüs keskendub esmajoones õppija keelelisele produktsioonile, sihtkeele produtseerimisprotsessi kohta saadud teave seevastu on juhuslikum või puudub üldse. Kuigi protsessi uurimiseks tuleb eelnevalt põhjalikult tunda produkti (vt Odlin 1989) ja produkti vahendusel on mingil määral võimalik uurida õppijate keeleloome- ja mõistmisprotsesse, ei anna õppija keeleline produktsioon üksi siiski piisavalt informatsiooni sihtkeele omandamise ja produtseerimise käsitlemiseks. Käesoleva artikli eesmärgiks ongi analüüsida probleeme, mis kaasnevad õppijakeele korpusanalüüsiga ning kirjeldada korpusanalüüsi täiendavaid introspektiivseid meetodeid, mis võimaldavad lisaks produktile süveneda ka õppija keeleloomeprotsessi.

2. Veaanalüüs teooria ja meetodina

Valdav osa õppijakeelt käsitlevatest uurimustest lähtub nii teoreetilise tausta kui ka meetodi poolest veaanalüüsist. **Teooriana** erineb veaanalüüs ja sellega tihedalt seotud vahekeele analüüs neile eelnenud strukturalistlik-kontrastiivsest teooriast selle poolest, et vigade ainsaks põhjustajaks ei peeta mitte enam õppija lähtekeelt, vaid vigade allikana on hakatud nägema ka sihtkeele süsteemi ja keeleõpetust (Sajavaara 1999b: 116). Veaanalüüsi ja vahekeele analüüsi teooria järgi osutavad vead keeleõppija probleemidele, mida iseloomustab süstemaatilisus.

Veaanalüüsi **meetodiks** on olnud vigade kogumine õppijate kirjalikest ja suulistest tekstidest. Läbi on viidud ka empiirilisi uurimusi veendumaks, kas õppijad tõepoolest teevad eeldatud vigu (Herranen 1978). Veaanalüüsis läbitakse tavaliselt järgmised etapid (James 1998): 1) vigade identifitseerimine,

mille käigus eristatakse kompetentsivead juhuslikest eksimustest; 2) vigade kirjeldamine; 3) vigade klassifitseerimine ja 4) vigade seletamine. Põhijoontes samadelt alustelt lähtub ka korpuspõhine veaanalüüs, mis erineb traditsioonilisest veaanalüüsist eelkõige suurema standardiseerituse ning normikohaste ja normivastaste kontekstide kõrvutamise poolest (Granger 2002: 11–16). Nii nagu traditsioonilises õppijakeele uurimises lähtub ka suurem osa õppijakeele korpuspõhiseid käsitlusi esmajoones veaanalüüsist. Seejuures on asjakohane rõhutada, et korpusanalüüsi kontekstis on veaanalüüs käsitletav meetodina, mitte aga teooriana (Hagen 2009). Viimasel ajal on siiski hakatud õppijakeelt korpuspõhiselt uurima muudeski teoreetilistes ja metodoloogilistes raamistikes (nt Nesselhauf 2003, 2005; Rasier & Hiligsmann 2007; Jantunen 2007; Eslon & Matsak 2009).

Kuigi õppijakeele veaanalüüs aitab paremini mõista õppijakeele olemust ja arengustaadiume ning omab olulisi pedagoogilisi rakendusi (Granger 2002: 14), on seda meetodit õigustatult kritiseeritud mitme metodoloogilise küsitavuse tõttu (vt Borin & Prütz 2004: 68; Ellis & Barkhuizen 2005: 70). Veaanalüüsile on ette heidetud lähteandmete heterogeensust, veakategooriate laialivalguvust, keeleomandamise käsitlemist staatilisena, piirdumist ainult sellega, mida õppijad teha ei suuda. Olulisemaid vajakajäämisi on väga juhuslik teave produtseerimisprotsessi ning õppijate strateegiate kohta. Samu probleeme ei ole suudetud vältida ka korpuspõhise veaanalüüsi juures. Järgnevalt leiavad käsitlemist mõned traditsioonilise ja korpuspõhise veaanalüüsi probleemid.

3. Korpuspõhise veaanalüüsi probleemid

Nii traditsioonilise kui ka korpuspõhise veaanalüüsi probleemid on eelkõige seotud vigade klassifitseerimise, vigade interpreteerimise ja vigade psühholingvistilise reaalsusega.

3.1. Vigade klassifitseerimine ja interpreteerimine

Õppijakeele korpuste veaklassifikatsioonid on reeglina lingvistilised: vigade klassifitseerimisel lähtutakse keelest kui paradigmaatiliste ja süntagmaatiliste seoste süsteemist. Paradigmaatilised seosed on hierarhilised: kõrgema tasandi keelend hõlmab madalamatasandilisi ja madalamatasandiline keelend sisaldub kõrgematasandilises. Süntagmaatilised seosed on lineaarsed, neis kombineerub leksikaalgrammatilise ja lausetasandi reeglistik. Keelesüsteemi süntagmaatilise ja paradigmaatilise telje ristumisel tekib kahemõõtmeline lingvistiline veaklassifikatsioon (Hufheisen 1991; Michiels 1999). Samast põhimõttest on lähtutud ka *Eesti vahekeele korpuse* (EVKK) veaklassifikatsiooni loomisel (lähemalt vt Eslon 2007: 106–109). Vealiike ja nende alamliike on EVKK veaklassifikatsioonis kokku 170, vealiikide määramise aluseks on ligi 300 tunnust.

Ükskõik kui detailne ei oleks mis tahes õppijakeele korpuse veaklassifikatsioon, ei ole see kunagi lünkadeta: klassifikatsiooni loomisel ei ole võimalik ette näha kõiki potentsiaalseid vealiike ja alamliike, mis õppijakeeles võivad esineda. Vajalik vealiik võib klassifikatsioonist puududa, nagu näites (1) esitatud lause puhul, milles on tegemist vale käändevormi kasutamisega ja ühildumisega ühildumatuse asemel:

(1) *Kõikedele **meile palvetele** ja meelepahadele nad vastasid ...*

Seetõttu on töö veaklassifikatsiooni täiustamisel pidev, uusi vealiike ja alamliike tuleb korpuse märgendamisel juurde, osa olemasolevatest võib aga osutuda mittevajalikuks.

Tõsiasi, et interpretatiivsus kuulub vea olemusse, on üldteada (vt Corder 1981). Tavapäraselt saab üks keeleviga ikka mitu märgendit ning see ei tulene veaklassifikatsiooni puudulikkusest, vaid asjaolust, et interpretatiivsus on veale iseloomulik. Näiteks võib õppija produtseeritud keelendit *proekteerimine* tõlgendada EVKK veaklassifikatsiooni alusel nii veana internatsionalismide kasutamisel (vealiik 1.6) kui ka häälduspärase kirjaviisina (vealiik 1.7). Vea mitmemõõtmelisel kirjeldamisel toetub EVKK märgendussüsteem veaklassi, -liigi ja alamliikide vahelistele seostele, samuti arvestatakse lähte- ja sihtkeele vaheliste sümmeetria, asümmeetria ja analoogiaseoste olemasoluga. Väidetavalt annab veaklasside, -liikide ja alamliikide väljatoomine ning alamliikide jagunemine omakorda kitsamateks vea ilminguteks võimaluse rääkida veast mitte ainult kui tagajärjest, vaid välja tuua ka vea tekkepõhjusi (vt Eslon 2006, 2007: 107). Tuleb siiski rõhutada, et vaid korpuspõhise veaanalüüsi alusel ning introspektiivsete meetodite abil õppija produtseerimisprotsessi uurimata saab rääkida vaid vigade oletatavatest, mitte aga tegelikest tekkepõhjustest. Samuti jääb pelgalt veaanalüüsi põhjal selgusetuks, milliseid keelendeid õppija veana tunnetab ja kas need vead on häirivad ka emakeelse keelekasutaja jaoks, teisisõnu, vigade psühholingvistiline reaalsus.

3.2. Vigade psühholingvistiline reaalsus

Keele omandamine ei ole kunagi lingvistiline probleem. Seda mõjutavad nii psühholoogilised kui ka sotsiaalsed tegurid, mis on tihedalt seotud lingvistiliste teguritega. Lingvistiliste, psühholoogiliste ja sotsiaalsete tegurite osakaal keele omandamisel ja omandamisprotsessi mõjutamisviis on eri tingimustes erinev ning kõiki lõpptulemust mõjutavaid tegureid veel ei tuntagi, kuigi sihtkeele omandamist on uuritud juba nelja aastakümne vältel (Sajavaara 1999a: 76). Seetõttu on vigade klassi-

fitseerimise ja interpreteerimise kõrval korpuspõhise veaanalüüsi olulisim probleem vigade psühholingvistiline reaalsus[5] (Kaivapalu 2004; Sajavaara 2006): kas ja kuidas viga tunnetatakse, kelle arvates ja mille suhtes on tegemist veaga.

EVKKs on keeleviga määratletud grammatikareeglile või kommunikatiivsele eesmärgile mittevastava keelekasutusena, mille hulka ei kuulu väsimusest ja hooletusest põhjustatud eksimused ning keelevääratused (Eslon 2007: 106). Grammatikareeglile või kommunikatiivsele eesmärgile vastavust ei tajuta aga alati sugugi üheselt, nagu ilmneb näites 2–4 esitatud sõnajärjevigade puhul:

(2) *3 aastat tagasi **ma lõpetasin** 10 klassi Narva Hummanitaar gümnaasiumis.*

(3) *Juba koolis **ma unistasin**, et tulevikus lähen TTÜ-sse õppima.*

(4) *Üks kord nädalas **tüdruk sõidab** linna, et käia kinos.*

Tegemist on struktuurilt sarnaste lausetega, milles kõigis on rikutud V2-reeglit: iseseisvas jaatavas väitlauses on öeldisverb (täpsemalt öeldisverbi pöördeline osa) tavaliselt teisel kohal, teemana toimiva moodustaja järel (EKK 2007: sy93.html). Hoolimata lausestruktuuri sarnasusest suhtusid erinevad märgendajad analoogilistesse sõnajärjevigadesse erinevalt: kui esimeses lauses märgendasid normivastase sõnajärje määrus–alus–öeldis sõnajärjeveaks kõik kolm märgendajat, siis teises lauses kaks ja kolmandas lauses vaid üks märgendaja. Selline tulemus on seda üllatavam, et lausetes (2) ja (3) on aluseks rõhutu asesõna, mille puhul peetakse otsejärge suhteliselt tavaliseks (EKK 2007: sy93.html). Seega tekib küsimus, millest sõltuvad

[5] Ligikaudu samas tähenduses on kasutatud ka terminit *psühhotüpoloogia* (vt Eckmann 2004).

märgenduseelistused, ehk teisisõnu, kas kirjakeele normile vastav V2-reegel on psühholingvistiliselt reaalne ka tegeliku keelekasutuse seisukohalt. Keelend, mis ei vasta kirjakeele normile, ei tarvitse tingimata olla vastuolus emakeelse keelekasutaja keeletajuga. Keelendi loomulikkus või ebaloomulikkus võib sõltuda vanusest, soost, haridusest jt taustateguritest, keeleõppija puhul lisanduvad võimalike keeletaju mõjutajatena veel lähtekeel ja sihtkeele oskuse tase.

4. Korpusanalüüsi täiendavaid meetodeid

Eelmises peatükis kirjeldatud veaanalüüsi probleemid on osaliselt lahendatavad introspektiivsete meetodite abil, mis täiendavad korpusanalüüsi nende uurimisküsimuste osas, mida viimane lahendada ei võimalda. Introspektiivsetest meetoditest on praktikas enim kasutatud psühholingvistilist tajutesti ja intervjuud, vähem valjusti mõtlemist ja õppija produtseerimisprotsessi fikseerivaid arvutiprogramme nagu ScriptLog. Järgnevalt kirjeldatakse lähemalt nende meetodite rakendusvõimalusi kombineerituna korpusanalüüsiga; põhjalikumalt peatutakse psühholingvistilisel tajutestil.

4.1. Psühholingvistiline tajutest

Lihtsaim võimalus uurida, milline keelend on emakeelse keelekasutaja või keeleõppija jaoks keeleomane ja loomulik, on küsida seda keelekasutajalt või -õppijalt endalt. Selleks kasutatakse ühe meetodina psühholingvistilisi tajuteste, mille käigus palutakse uurimuses osalejatel hinnata keelendi loomulikkust Likert-tüüpi skaalal *täiesti loomulik – pigem loomulik – pigem ebaloomulik – täiesti ebaloomulik.* Seejuures rõhutatakse, et lähtuda tuleb oma keeletundest, mitte aga keelendite normikohasusest.

Kuna EVKK sagedaseim veatüüp on sõnajärjevead, mille märgendamisel märgendajate hinnangud sõnajärjemalli normikohasusele ja normivastasusele sageli erinevad, siis kasutati sõnajärjemallide loomulikkuse uurimiseks psühholingvistilist tajutesti (vt lisa 1). Testilausete aluseks oli 14 sõnajärjevea tüüpi, mille osas märgendajad olid üksmeelsed, näiteks

(5) *Aasta jooksul **nad kasvatavad** põrsast.*

Õppija vigasest lausest moodustati „Eesti keele käsiraamatu" (2007) alusel normikohane lause

(6) *Aasta jooksul **kasvatavad nad** põrsast.*

Normikohase lause (6) sõnajärge muudeti nii lause moodustajate ümberpaigutamise teel kui ka sõnavara varieerimise teel. Igast sõnajärjevea tüübist moodustati sel viisil 3–5 varianti, mille normikohasus võis olla varieeruv, nagu näiteks lause (6) variandid (7) – (9):

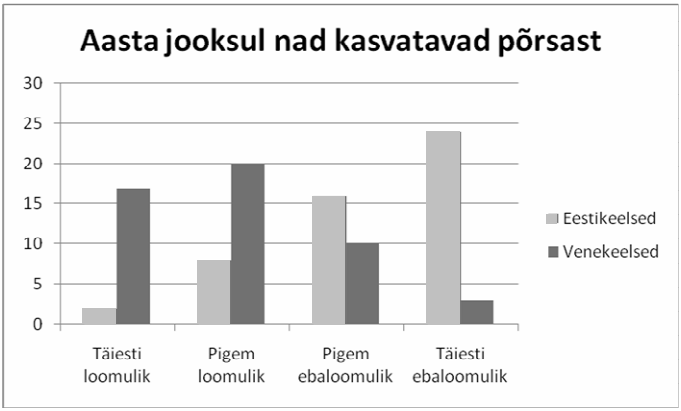
(7) *Nad õpivad terve nädala matemaatikat.*

(8) *Pool päeva lugesime me kriminulle.*

(9) *Luuletust kirjutas ta päev läbi.*

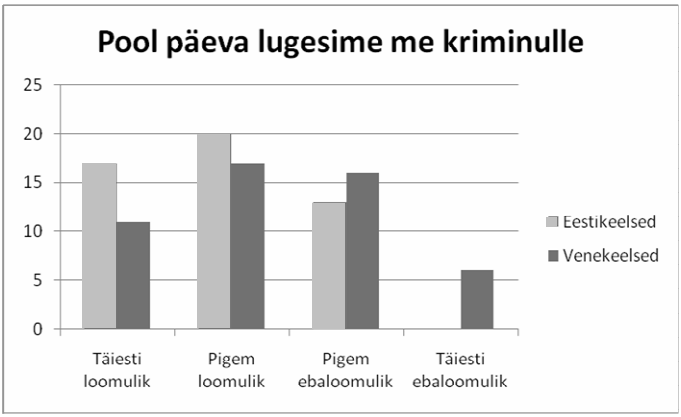
Uurimuses osalesid Tallinna Ülikooli üliõpilased vanuses 19–37 aastat. Osalejate hulgas oli 50 eestikeelset keelekasutajat ja 50 venekeelset eesti keele õppijat, kes olid koolis õppinud eesti keelt 6–12 aastat. Uurimuses ei kasutatud nende informantide vastuseid, kes olid õppinud eestikeelses koolis, kuid kelle emakeel oli vene keel ja kodune keel eesti keel või kelle emakeel oli eesti keel ja kodune keel vene keel. Uurimuse tulemused on esitatud joonistel 1–4.

Testivastustest selgus, et V2-reeglit eiravat sõnajärge määrus–alus–öeldis (vt joonis 1) tajusid täiesti ebaloomulikuna või pigem ebaloomulikuna 80% eestikeelsetest keelekasutajatest. Seevastu enamiku venekeelsete keeleõppijate jaoks (37 informanti 50-st) oli selline sõnajärg täiesti loomulik või pigem loomulik.



Joonis 1. Eestikeelsete keelekasutajate ja venekeelsete eesti keele õppijate hinnang lause *Aasta jooksul nad kasvatavad põrsast* loomulikkusele

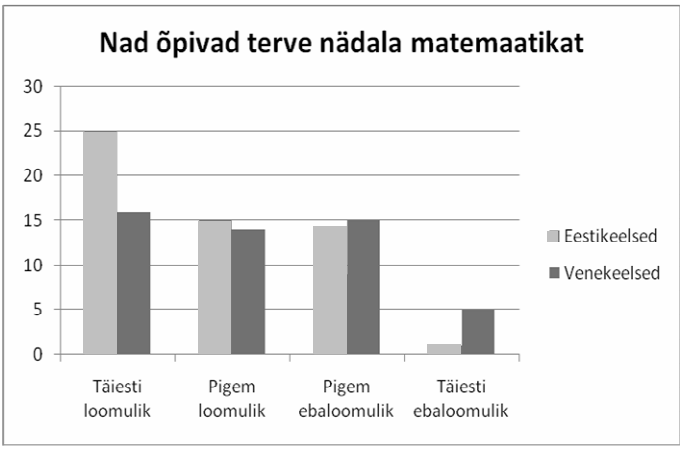
V2-reeglit järgivat sõnajärge (vt joonis 2) tajus täiesti loomulikuna või pigem loomulikuna 37 informanti 50 eestikeelsest keelekasutajast. Asjaolu, et 13 eestikeelse keelekasutaja jaoks



Joonis 2. Eestikeelsete keelekasutajate ja venekeelsete eesti keele õppijate hinnang lause *Pool päeva lugesime me kriminulle* loomulikkusele

tundus V2-sõnajärg olevat pigem ebaloomulik, võib muuhulgas olla tingitud mitmuse esimese pöörde lõpu ja personaalpronoomeni kõrvuti asetsemisest. Viimane oletus vajaks aga täpsustamist intervjuu või valjusti mõtlemise abil. Ka pisut enam kui poolte venekeelsete õppijate jaoks (28 informanti 50-st) oli antud sõnajärg täiesti või pigem loomulik.

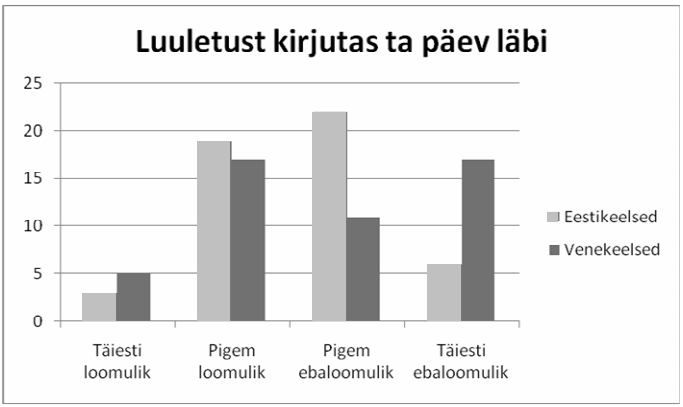
Tavapärane normaallause sõnajärjemall alus–öeldis–määrus–sihitis (vt joonis 3) oli täiesti loomulik või pigem loomulik 80% eestikeelsete keelekasutajate ning 60% venekeelsete keeleõppijate jaoks. 40% keeleõppijatest tajus antud sõnajärge siiski pigem või täiesti ebaloomulikuna.



Joonis 3. Eestikeelsete keelekasutajate ja venekeelsete eesti keele õppijate hinnang lause *Nad õpivad terve nädala matemaatikat* loomulikkusele

Järgmises lauses (joonis 4) on tegemist nn fokuseeritud sõnajärjega: rõhutatav lauseliige on paigutatud lause algusesse markeerimaks, et just luuletust, aga mitte romaani, novelli või näidendit kirjutati päev läbi. Testivastustest selgus, et kon-

tekstivaba fokuseeritud sõnajärjega lauset tajusid nii eestikeelsed keelekasutajad kui ka keeleõppijad ühtviisi pigem ebaloomuliku kui loomulikuna: 28 keelekasutaja ja keeleõppija jaoks oli antud sõnajärg pigem või täiesti ebaloomulik.



Joonis 4. Eestikeelsete keelekasutajate ja venekeelsete eesti keele õppijate hinnang lause *Luuletust kirjutas ta päev läbi* loomulikkusele

Psühholingvistiline tajutest näitas niisiis V2-reegli psühholingvistilist reaalsust eestikeelse keelekasutaja jaoks: V2-reegli rikkumist tajuti pigem ebaloomuliku kui loomulikuna. Seevastu venekeelsete keeleõppijate jaoks on V2-reegli rikkumine pigem loomulik kui ebaloomulik. Samas tajusid pisut enam kui pooled keeleõppijatest V2-reeglile vastavat sõnajärge täiesti loomulikuna või pigem loomulikuna. Seega ei ole V2-reegli järgimine või rikkumine venekeelse õppija keeletaju seisukohalt distinktiivne: loomulikuna tajuti nii kirjakeele normi kohast sõnajärge määrus–öeldis–alus kui ka normivastast sõnajärge määrus–alus–öeldis. Eestikeelse keelekasutaja jaoks väljendab aga V2-reegli järgimine loomulikku ning selle rikku-

mine ebaloomulikku keelekasutust. Järelikult on põhjendatud V2-reegli rikkumise märgendamine sõnajärjeveana: emakeelse keelekasutaja keeletaju toetas antud juhul õigekeelsusreeglit ning samas ilmnes, et keeleõppija ei taju selgelt nende sõnajärjemallide erinevust.

Kirjeldatud psühholingvistiline tajutest võimaldab uurida, milline keelend on keelekasutaja ja/või -õppija jaoks loomulik ja milline mitte, ehk teisisõnu, kaardistab olukorda. Samas ei anna tajutest teavet selle kohta, miks keeleõppija keelendit loomulikuna või ebaloomulikuna tajub, millised tegurid selle tajumist põhjustavad. Keelekasutaja ja/või -õppija keeletaju põhjuste väljaselgitamiseks sobivad valjusti mõtlemine ja intervjuu, mida kasutatakse sageli kombineerituna (nt Kaivapalu 2005).

4.2. Valjusti mõtlemine

Valjusti mõtlemine (ingl *thinking aloud*, sm *ääneenajattelu*) on meetod, mille eesmärgiks on pääseda ligi mentaalsetele protsessidele, mille kohta muul viisil informatsiooni ei saa (Swain, Lapkin 1995). Valjusti mõtlemine seisneb selles, et keelekasutaja või -õppija kommenteerib testi sooritamise või ülesande täitmise ajal valjusti oma tegevust ja põhjendab oma tegutsemisviisi. Kommentaarid lindistatakse ja litereeritakse. Valjusti mõtlemine võimaldab uurida keeleloome- ja mõistmisprotsessi mehhanisme ning iseärasusi. Meetodit on kritiseeritud, väites, et tegemist on kunstliku olukorraga, mitte spontaanse keelekasutusega, samas aga ei ole paremat lahendust mentaalsete protsesside uurimiseks pakutud. Enam on valjusti mõtlemist kasutatud teksti mõistmise uurimiseks (vt Block 1992), kuid see on andnud väga häid tulemusi ka õppijate produtseerimisprotsessi iseärasuste väljaselgitamisel (vt Kaivapalu 2005). Näited (10) ja (11) kirjeldavad eestikeelsete soome keele

õppijate kommentaare oma käänamisprotsessi kohta (Kaivapalu 2005: 269–270):

(10) *toinen toisia,* see võiks nüüd olla selle järgi, et eesti keeles on *teine* ja *teisi.* Sisseütlev on *toisiin* ja seestütlev on *toisista*

(11) *pankki* See on teistsugune kui *herkku* ja *kirkko. pankkeja pankkeja pank/kei/ta* ei *pan- pankkei/siin pan- pankei- pan/koista* või *pan/keista pankeista* siiski.

Esimene kommentaar annab tunnistust lähtekeele (eesti keel) analoogia kasutamisest sihtkeele (soome keel) mitmuse partitiivi produtseerimisel, teises näites püüab õppija erineva tüvevokaaliga sihtkeele sõnu omavahel võrreldes ja analoogiaseost otsides leida õiget käändeparadigmat.

Valjusti mõtlemine sobib korpusanalüüsi täiendama eelkõige heitmaks valgust keelekasutaja ja -õppija kirjutamisprotsessile: kuidas sõnu valitakse, käände- ja pöördevorme produtseeritakse, lauseid moodustatakse ja tekstiks seotakse. Samuti sobib meetod suurepäraselt teksti mõistmise uurimiseks. Seejuures tuleb arvestada, et kõik informandid ei ole ühtviisi altid oma tegevust kommenteerima, samuti ei suuda kõik inimesed ühtviisi hästi keskenduda samaaegselt teksti kirjutamisele ja sellest rääkimisele. Valjusti mõtlemise kui uurimismeetodi kasutamise tulemusel selgusetuks jäänud küsimusi võib täpsustada retrospektiivse intervjuu abil.

4.3. Retrospektiivne intervjuu

Retrospektiivne intervjuu viiakse läbi vahetult valjusti mõtlemise järel, et intervjueerija ei jõuaks unustada küsimusi, mida ta tahab täpsustada ja süvitsi uurida, ega intervjueeritav oma tegutsemise põhjendusi. Retrospektiivsel intervjuul nagu igal teist tüüpi intervjuulgi on meetodina mitmeid eeliseid (vt Hirsjärvi & Remes & Sajavaara 2005: 191–194; Hirsjärvi & Hurme

2000): intervjueerija on intervjueeritavaga vahetus keelelises interaktsioonis; ainestiku kogumine on paindlik; intervjueeritav saab ennast vabalt väljendada ning ta on tähendusi loov ja seega aktiivne osapool. Retrospektiivse intervjuu küsimused kujunevad otseselt uurimuses osalejate testi sooritamise või ülesande lahendamise ja valjusti mõtlemise põhjal. Nii on näites (12) täpsustatud, kas õppija toetub sihtkeele sõnavormide produtseerimisel lähtekeelele või mitte (Kaivapalu 2005: 210):

(12) A.K.: *Kui sa mõtled, et sõna on nagu eesti keeles, kas sa siis moodustad eesti keele järgi ka või?*

R.: *Ei, eesti keele järgi ma ei moodusta.*

A.K.: *Miks sa ei moodusta, oskad sa öelda?*

R.: *Ma ei ole hakand nagu moodustama, proovin ikka mõelda, kuidas soome keeles on. Need on ikka üpris erinevad ka, eesti keeles moodustamine ja soome keeles moodustamine. Mulle on jäänud selline mulje, siis ei hakka.*

A.K.: *Kardad, et ajab segadusse?*

R.: *Nojah.*

A.K.: *Ei julge nagu eesti keelt appi võtta?*

R.: *Ei.*

Nii nagu valjusti mõtlemise puhul, on ka retrospektiivse intervjuu eesmärgiks korpusanalüüsi täiendamine õppija produtseerimisprotsessi käsitleva teabega.

4.4. Keeletehnoloogilised vahendid: arvutiprogramm ScriptLog

Produtseerimisprotsessi uurimiseks on loodud ka keeletehnoloogilisi vahendeid. Enim on kasutamist leidnud arvutiprogramm ScriptLog[6]. Tegemist on programmiga, mis fikseerib

[6] http://www.scriptlog.net (19.8.2009).

kirjutamisprotsessi: salvestab õppija peatumised, tagasipöördumised ja parandused. Kuna igasugune peatumine produtseerimisel annab tunnistust õppija probleemidest, vajadusest järele mõelda, siis sisaldavad peatumised, tagasipöördumised ja parandused väärtuslikku teavet selle kohta, milliste raskustega õppija tekstiloomes kokku puutub ning kuidas ta neid lahendab. Nii on Scriptlogi kasutatud näiteks poolakeelsete õppijate, kelle teine keel on saksa keel, rootsi keele kirjutamisprotsessi (vt Kowal 2009) uurimiseks, et välja selgitada õppijate esimese ja teise keele vaheline tööjaotus kolmanda keele omandamisel. Uurimus näitas, et teise keele (saksa keele) interferents rootsi keele omandamisele ilmnes eelkõige sõnatasandil (sõnavalik, sõnamuutmine) ning esimese keele (poola keele) interferents süntaktilisel ja tekstitasandil. Nii nagu psühholingvistiliste tajutestide puhulgi, ei anna programm informatsiooni probleemide põhjuste kohta ning need tuleb välja selgitada valjusti mõtlemise meetodil või intervjuude abil.

Kuigi teadaolevalt ei ole ScriptLogi korpusainestiku kogumisel seni kasutatud, on sellel kirjutamisprotsessi uurimise seisukohalt ulatuslik perspektiiv. Seetõttu on kavas luua EVKKsse ScriptLogi allkorpus.

5. Kokkuvõtteks

Korpusanalüüs on õppijakeele universaalsete ja keelespetsiifiliste joonte kaardistamisel ning õppijakeeles esinevate suundumuste väljaselgitamisel asendamatu eelkõige tulemuste kõrge üldistusastme tõttu. Seda võimaldab ulatuslik andmestik. Õppijakeele nähtuste olemuse ja põhjuste süvitsi uurimine eeldab aga korpusanalüüsi täiendamist introspektiivsete meetoditega, millest olulisemad on valjusti mõtlemine ja sellele järgnev retrospektiivne intervjuu testi sooritajaga. Samamoodi

on võimalik kombineerida teksti kirjutamist produtseerimisprotsessi fikseeriva arvutiprogrammi ScriptLog abil ja retrospektiivset intervjuud. Oluline on niisiis ainestiku triangulatsioon: ainult ühe meetodi kasutamine piirab oluliselt õppijakeele uurimisvõimalusi ega luba teha põhjapanevaid järeldusi keele omandamise universaalsete ja keelespetsiifiliste joonte kohta. Korpuspõhist veaanalüüsi tuleks õppijakeele uurimisel käsitleda vaid ühe meetodina, millest lähtutakse õppijakeele probleemsete piirkondade kaardistamisel ja mida saab rakendada samuti keeletehnoloogiliste lahenduste leidmisel. Veaanalüüsi puhul ei tuleks kindlasti arvestada ainult kirjakeele normi, vaid veenduda ka normivastase keelendi psühholingvistilises reaalsuses nii emakeelse keelekasutaja kui ka keeleõppija seisukohalt.

Kirjandus

Block, Ellen L. 1992. See how they read: Comprehension monitoring of L1 and L2 readers. − TESOL Quarterly 26 (2), 319–343.

Borin, Lars & Prütz, Klas 2004. New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. – Corpora and Language Learners / Ed. by G. Aston & S. Berardini & D. Stewart. Amsterdam / Philadelphia: John Benjamin Publ. Co, 67–87.

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

EKK 2007 = Eesti keele käsiraamat. http://www.eki.ee/books/ekk07/ (19.8.2009).

Eckman, Fred 2004. From phonemic differences to constrant rankings: Research on second language phonology. − Studies in Second Language Acquisition 26, 514–539.

Ellis, Rod & Barkhuizen, Gary 2005. Analyzing learner language. Oxford University Press.

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

Eslon, Pille 2007. Õppijakeelekorpused ja keeleõpe. − Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine. Tallinna Ülikooli eesti filoloogia osakonna toimetised 9 / Toim P. Eslon. Tallinn: TLÜ Kirjastus, 87−120.

Eslon, Pille & Matsak, Erika 2009. Eesti keele kasutusvariandid: korpusest tulenev käändevormide võrdlev analüüs. – Eesti Rakenduslingvistika Ühingu aastaraamat 5 / Toim. H. Metslang & M. Langemets & M.-M. Sepper, R. Argus. Tallinn: Eesti Keele Sihtasutus, 79–110.

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.

Granger, Sylviane 2002. A bird's eye view of learner corpus research. – Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching / Ed by S. Granger & J. Hung & S. Petch-Tyson. Amsterdam / Philadelphia: John Benjamin Publ. Co, 3–33.

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 (20.6.2009).

Hagen, Jon Erik 2009. Strukturfeil og kateogorfeil i skriftlig andrespråksperformanse. Den 9. nordiske andetsprogskonference. Helsingør, 11.–13. juni 2009.

Herranen, Tauno 1978. Errors made by university students in the use of the English article system. – Further contrastive papers. Jyväskylä Contrastive Studies 6 / Toim. K. Sajavaara & J. Lehtonen & R. Markkanen. Jyväskylä: University of Jyväskylä, 74–95.

Hirsjärvi, Sirkka & Hurme, Helena 2000. Tutkimushaastattelu. Teemahaastattelun teoria ja käytäntö. Helsinki: Yliopistopaino.

Hirsjärvi, Sirkka & Remes, Pirkko & Sajavaara, Paula 2005. Uuri ja kirjuta. Tallinn: Medicina.

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

James, Carl 1998. Errors in language learning and use. Exploring error analysis. London: Longman.

Jantunen, Jarmo 2007. Oppijansuomen piirteitä korpusvetoisesti. – VIRSU III. Suomalais-ugrilaisia kohdekieliä ja kontakteja. Studies in Languages 42 / Toim. P. Muikku-Werner & O. Kokko & H. Remes. Joensuu: Joensuun yliopisto, 69−84.

Kaivapalu, Annekatrin 2004. Kui sarnane on sarnane? Eesti ja soome mitmusevormide psühholingvistilisest reaalsusest. – VIRSU II. Suomi ja viro kohdekielinä. Oulun yliopiston suomen ja saamen kielen ja logopedian laitoksen julkaisuja 24 / Toim. H. Sulkala & H. Laanekask. Oulu, 62–71.

Kaivapalu, Annekatrin 2005. Lähdekieli kielenoppimisen apuna. Jyväskylä Studies in Humanities 44. Jyväskylä: Jyväskylän yliopisto.

Kowal, Iwona 2009. Svenska som L3. Var döljer sig L1 och L2? Den 9. nordiske andetsprogskonference. Helsingør, 11.–13. juni 2009.

Kaivapalu, Annekatrin 2008. Lähtekeele mõju korpuspõhine uurimine. – Õppijakeele analüüs: võimalused, probleemid, vajadused. Tallinna Ülikooli eesti filoloogia osakonna toimetised 10 / Toim P. Eslon. Tallinn: TLÜ Kirjastus, 93−119.

Michiels, B. 1999. Die Rolle der Niederländischkenntnisse bei Französischsprachigen Lernern von Deutsch als L3: Eine empirische Untersuchung. – Zeitschrift für Interkulturellen Fremdsprachunterricht 3(3). http://www.spz.tu-darmstadt.de/projekt_ejournal/jg-03-3/beitrag/mich1.htm (18.6.2009).

Nesselhauf, Nadja 2003. The use of collocations by advanced learners of English and some implications for teaching. − Applied Linguistics 24: 2, 223−242.

Nesselhauf, Nadja 2005. Collocations in learner corpus. Amsterdam: John Benjamins Publishing Company.

Odlin, Terrence 1989. Language Transfer. Cross-linguistic influence in language learning. Cambridge, USA: Cambridge University Press.

Rasier, Laurent & Hiligsmann, Philippe 2007. Prosodic transfer from L1 to L2. Theoretical and methodological issues. − Nouveaux cahiers de linguistique française 28, 41−66.

Sajavaara, K. 1999a. Toisen kielen oppiminen. − Kielenoppimisen kysymyksiä. Soveltavan kielentutkimuksen keskus / Toim. K. Sajavaara & A. Piirainen-Marsh. Jyväskylä: Jyväskylän yliopisto, 73–102.

Sajavaara, K. 1999b. Kontrastiivinen kielentutkimus ja virheanalyysi. − Kielenoppimisen kysymyksiä. Soveltavan kielentutkimuksen keskus / Toim. K. Sajavaara & A. Piirainen-Marsh. Jyväskylä: Jyväskylän yliopisto,103–128.

Sajavaara, Kari 2006. Kontrastiivinen analyysi, transfer ja toisen kielen oppiminen. – Lähivertailuja 17. Jyväskylä Studies in Humanities 53 / Toim. A. Kaivapalu & K. Pruuli. Jyväskylä: Jyväskylän yliopisto, 9−26.

Swain, Merrill & Lapkin, Sharon. 1995. Problems in output and the cognitive processes they generate: A step toward second language learning. − Applied Linguistics 16 (3), 371−391.

Tarnanen, Mirja 2007. Testiaineistosta kielenoppijakorpukseksi. – Kieli oppimisessa – Language in Learning. AFinLAn vuosikirja 65 / Toim. O.-P. Salo & T. Nikula & P. Kalaja. Jyväskylä: Jyväskylän yliopisto, 197−214.

Tenfjord, Kari 2008. ASKeladden – a corpus-based approach to L1 transfer in Norwegian learner language. http://www.askeladden.uib.no/summary.page (19.8.2009).

LISA 1.
Test eesti keele sõnajärjemallide loomulikkuse uurimiseks

Sõnajärje loomulikkuse uuring

Eesti keele sõnajärge loetakse võrreldes mõne teise keelega suhteliselt vabaks. Siiski tunduvad mõned sõnajärjemallid nii emakeelsele keelekasutajale kui ka keeleõppijale loomulikumad kui teised. Käesoleva uuringu eesmärgiks on uurida eesti keele teatud sõnajärjemallide loomulikkust. Seepärast ärge mõelge, kas Teie vastused on õigekeelsuse seisukohast õiged või valed, vaid vastake nii, nagu Teie enda keeletunne Teile ütleb. Täname Teid väga uuringule kaasaaitamise eest!

I. Taustandmed vastaja kohta

Vanus

.....

Sugu

.....

Emakeel

.....

Kodune keel

.....

Kus ja kui kaua eesti keelt õppinud

.....

.....

II. Sõnajärjemallide loomulikkuse taju

Palun hinnake järgnevaid lauseid oma keeletundest lähtuvalt ja tehke rist sellesse lahtrisse, mis vastab kõige täpsemini Teie keeletundele.

	Lause	Täiesti loomulik	Pigem loomulik	Pigem ebaloomulik	Täiesti ebaloomulik
1	*Nad õpivad terve nädala matemaatikat.*				
2	*Teemat muudame me vabalt.*				
3	*Põhiteema oli elu planeedil Marss tõenäosus.*				
4	*Siidist õmmeldud kleidi kinkis ema tütrele.*				
5	*On vaja õpilastele raamatuid.*				
6	*Järelikult su teod paistavad praegu välja kentsakad.*				
7	*Teatud puudused on isegi pangas töötamisel.*				
8	*See on ainuke ülikool, mis üldse välja näeb nagu ülikool.*				
9	*Ta kirjutas lõpukirjandi paari aasta eest Tallinnas.*				
10	*Pärast seda ma kolisin Peterburi.*				
11	*Nad põnevalt jutustasid oma õppeainet.*				

12	*Ma saan uue võimaluse alles pärast vaheaega.*				
13	*Aasta jooksul nad kasvatavad põrsast.*				
14	*Vivian teeb kodutöid sageli rutakalt.*				
15	*Vestlusteema oli eksperimendi olulisus uuringus CEFR.*				
16	*Mind aitab see eriti tulevases töös.*				
17	*Esimest korda toimunud aastal 1869 laulupidu muutus ajaloolise tähtsusega sündmuseks.*				
18	*Zaura kinkis mulle siili tehtud paberist.*				
19	*Lastele oli tarvis vanemaid.*				
20	*Elamisel õpilaskodus on ka oma miinused.*				
21	*Mõne aasta pärast alustavad nad magistriõpinguid välismaal.*				
22	*Mõtles poiss pingsalt oma isast.*				
23	*Pärast kooli ma sain hea keeleoskuse.*				
24	*Aastal 2003 esimest korda ilmunud monograafia sai üldise tunnustuse osaliseks.*				

25	*See on ainus küsimus, mis tegelikult kerkib üles nagu probleem.*				
26	*Pool päeva lugesime me kriminulle.*				
27	*Ta alati riietub pidulikult.*				
28	*Olemise võimalikkus allvee-laeval Nautilus on probleem.*				
29	*Mu elu praegu näeb välja järgmiselt.*				
30	*Nad lendasid mõne aja pärast Indiasse.*				
31	*Tehnikumi järel sai ta hea töökoha.*				
32	*Poiss kinkis tüdrukule puust meisterdatud kujukese.*				
33	*Energiat on vaja inimestele.*				
34	*See on rikkalik kingitus, mis kuldab mind täiesti üle nagu miljonäri.*				
35	*See väga aitas mind järgmises koolituses.*				
36	*Ta rääkis huvitavalt lossi ajaloost.*				
37	*Kutsekoolis õppimisel on samuti oma head küljed.*				
38	*Mõtlesid tüdrukud pärast trenni minna raamatukokku.*				
39	*Kergelt vahetavad nad värvi.*				

40	*Õpetajad kirjutasid ministeeriumi pärast seaduse vastuvõtmist.*				
41	*Aastal 2002 toimunud konverents muutus esimest korda tähelepanuväärseks sündmuseks.*				
42	*Edu ainult saatis teda kõikjal.*				
43	*Ülikooli lõpetamise järel sõitis Henrik Londonisse.*				
44	*Edasises elus teid juhib vaid südametunnistus.*				
45	*Nüüd läheb meie olukord ülemustele tõsiselt korda.*				
46	*Vanad raamatud saame üle anda pärast aastavahetust.*				
47	*Rahva tahtest kirjutas innustavalt ajaleht.*				
48	*Kolm aastat tagasi ma lõpetasin kümnenda klassi Narva Humanitaargümnaasiumis.*				
49	*Luuletust kirjutas ta päev läbi.*				
50	*Ajaloolise tähtsusega sündmuseks kujunes esimest korda aastal 1900 läbi viidud rahvaloendus.*				

On some methods of contributing to corpus-based analysis of learner language

Annekatrin Kaivapalu

Summary

Most of corpus-based studies of learner language have been completed in the framework of error analysis and have focused mainly on the learners' products, while learners' production processes have been less researched. This article deals with some problems and limitations of corpus-based error analysis of learner language. The problems of classification and interpretation as well as the psycholinguistic reality of errors are discussed on the bases of word order errors of Russian learners of Estonian. Some complementary possibilities enabling the researcher to investigate learners' production process are proposed: introspective methods such as psycholinguistic perceptual tests, *thinking aloud*-method, retrospective interviews and language technology software such as *Script-Log*-program.

The article also presents some results of a preliminary study on psycholinguistic reality of some word order patterns of Estonian from the point of view of Russian learners of Estonian as well as from that of native Estonian speakers. The following word order patterns are discussed: 1) adverbial-subject-predicate-object, 2) adverbial-predicate-subject-object, 3) subject-object-adverbial-object, 4) object-predicate-subject-adverbial. The study concludes that the main differences in perception of the Estonian native speakers and the Russian learners of Estonian concerned V2-rule: the pattern adverbial-predicate-subject-object was perceived as natural by most of

Estonian native speakers and the pattern adverbial-subject-predicate-object was mostly perceived as unnatural. For most of the Russian learners the pattern adverbial-subject-predicate-object was quite natural or nearly natural. The pattern adverbial-predicate-subject-object was perceived as quite natural or nearly natural by a half of the Russian learners of Estonian.

Keywords: corpus-based error analysis, psycholinguistic reality of errors, word order patterns of Estonian, Estonian native speaker, Russian learner of Estonian, introspective methods

Autor

PhD Annekatrin Kaivapalu, Tallinna Ülikooli eesti keele ja kultuuri instituudi dotsent, riikliku „Eesti keele keeletehnoloogiline tugi (2006–2010)" projekti „VAKO – Eesti vahekeele korpuse keeletarkvarana ja keeletehnoloogilise ressursi arendamine" põhitäitja, kaivapa@tlu.ee

KOMAVIGADE TUVASTAJA

Krista Liin

Ülevaade

Artiklis tutvustatakse eesti keelele komavigade kontrolliks loodud grammatikakorrektori prototüüpi. Kitsenduste grammatika formalismis kirjutatud reeglipõhine programm tegeleb tekstis üleliigsete ning puuduvate komade tuvastamisega, kuid ei tee veel parandusi. Märgendatakse sõnaliike ja -vorme, mille ees sageli komadega eksitakse: sidesõnu, küsisõnu ja verbide pöördelisi vorme. Reeglid põhinevad grammatikaõpikuis leiduvatel õigekirjareeglitel, lisaks on reeglite loomisel ja testimisel kasutatud kolme eri tekstiliiki kuuluvaid korpustest leitud tekstinäiteid. Peamine korpus koosneb internetiportaalide kommentaaridest kogutud väära komakasutusega lausetest, lisaks on komavigade tuvastajat testitud ka ajalehe- ja teadustekstidel. Artiklis antakse ülevaade vigade tuvastamisel tekkinud probleemidest ning nende võimalikest põhjustest. Saavutatud tulemused on võrreldavad Põhjamaade teiste grammatikakorrektoritega.

Võtmesõnad: automaatne veatuvastus, grammatikakontroll, komavead[1]

[1] Komavigade tuvastaja loomist on toetanud riikliku programmi "Eesti keele keeletehnoloogiline tugi" (2006–2010) projekt "Süntaksianalüüsil põhinev keeletarkvara ning selle arendamiseks vajalikud keeleressursid" (2006–2008).

1. Grammatikakorrektori ülesanded

Kirjutaja abivahenditest on eesti keele kättesaadav speller ehk õigekirjakorrektor, mis tuvastab kirjutamise käigus tekkivad ortograafiavead. Samas toetub speller piiratud lingvistilisele infole, vaadates reeglina korraga üksnes üht sõna ning vastava sõnavormi esinemist keeles. Seega jäävad spelleri poolt märkimata need juhud, kus on küll tehtud õigekirjaviga, kuid saadud mõni teine, konteksti sobimatu, kuid eesti keeles võimalik sõnavorm (näiteks *kindlasti* asemel on kirjutatud *kindalasti*). Abivahendit, mis arvestaks õigekeelsuse üle rohkema, morfoloogilise ja süntaktilise info põhjal kasutaks otsustamisel suuremat konteksti kui üksainus sõna ning kontrolliks siis ühe lause või mitmest lausest koosneva tekstiosa sobivust, nimetatakse grammatikakorrektoriks.

Grammatikakorrektor kasutab oma otsustustes nii morfoloogilisi kui ka süntaktilisi teadmisi, võimaldades tuvastada õigekirja-, ühilduvus-, kokku-lahkukirjutamise ja kirjavahemärgivigu, parandada sõnavalikut ja sõnajärge ning muudki. Kirjutaja seisukohalt ei piisa tekstiredaktorist, mis üksnes vigased kohad ära märgib. Õigekirjakontroll peaks võimaldama ka (poolautomaatset) parandust: pakkuma iga vea kohta vähemalt üht varianti, kuidas korrektne lause välja näeks. Samas on mõttekas hoida parandused vigade tuvastusest eraldi, kuna alati ei pruugi kohene parandamine soovitav olla. Näiteks on võimalik kontrollida lausete õigekirja keeleõppeprogrammides, kus õpilasele ei näidata kohe kogu infot, vaid antakse kõigepealt teada vea liigist ja kohast lauses ning alles teatud tingimustel näidatakse ülesande lahendust – korrektset lauset. Teisest küljest võib osutuda vajalikuks ka kohene automaatne korrektuur. Seda näiteks juhul, kui grammatikakorrektor on osa pikemast automaatsest protsessist ning järgmine samm, olgu selleks siis süntaksianalüüs, vajab vigadeta sisen-

dit. Niisiis oleks kasulik ehitada grammatikakorrektor kaheosalisena: vigade märgendamise osa, mida saab soovi korral kasutada koos teise, vigade korrigeerimist võimaldava osaga.

2. Komavigade tuvastamine

Sageli alustatakse grammatikakorrektori loomist kas ühilduvus- või interpunktuatsioonivigade märgendamisest. Enamasti valitakse välja üks nimetatud vealiikidest ning jäetakse teise veatüübi kontroll hilisemaks. Seda tehakse tihti arvestusega, et tolle sisendis on esimest liiki vead juba parandatud. Arvamused selle kohta, kas enne tuleks tegelda kirjavahemärkide või ühilduvusega, lähevad lahku. Eckhard Bick (2006: 9) argumenteerib taani grammatikakorrektori kirjelduses, et komavigu tuleks tuvastada grammatilisemas kontekstis, kus lause on juba muus osas keeleliselt korrektne. Teisest küljest toovad baski keele grammatikakorrektori loojad välja, et muid ühildumisvigu on tunduvalt kergem tuvastada ja parandada, kui komad ja seega ka osalausepiirid on korrektselt määratud (vt Aldezabal jt 2003: 1). Niisiis pole selget üksmeelt, millist veatüüpi tuleks kõigepealt korrigeerida, kuna iga vea vähendamine aitab pea alati kaasa järgmiste paremale tuvastamisele.

Eesti keele grammatikakorrektori puhul on valik langenud komavigade tuvastamisele. Eksimused kirjavahemärkide kasutuses on muudest vealiikidest võrdlemisi kergesti eristatavad, samuti võiks eeldada, et kuna kirjavahemärke kasutatakse üksnes tekstis, mitte suulises kõnes, siis sõltuvad selle valdkonna vead suhteliselt vähe muus osas keeleliselt korrektse lause moodustamisest. Teisisõnu, komavigu teevad kirjas ka need, kes on suulise keele suurepäraselt omandanud. Teine põhjus komavigade valikuks oli asjaolu, et teiste keeltega võrreldes tundub komakasutus eesti keeles olevat suhte-

liselt kindlalt reguleeritud, samas aga suhteliselt keeruline ja raskesti omandatav. Seega on olemas praktiline vajadus komavigade automaatse tuvastamise järele.

Komakasutuses on võimalik eristada kaht tüüpi vigu: koma ärajätmine seal, kus see olema peaks (puuduv koma), ja koma asetamine sinna, kus seda vaja pole (üleliigne koma). Vaadates üldist keelekasutust, võib öelda, et pigem jätame koma lausest välja, olgu selle põhjuseks siis hooletus või eeldus, et paus lauses mõne muu vahendiga esile tuuakse. Nii on juhtunud ka näites (1), kus paksus kirjas olevate sõnade ees peaks koma olema.

(1) *Müüsin vana auto maha hetkel **mil** oleks pidanud riskigrupp langema 0.77-0.60-le **aga** uut autot kindlustama minnes tõusis hoopis 0.87.*

Kuigi valitseb tendents koma lausest välja jätta, on ka juhtumeid, kui kas segadusseajavatest õigekirjareeglitest juhindudes või mõnel muul põhjusel kirjutatakse koma sinna, kus see reeglite kohaselt olema ei peaks ega olla tohiks. Sel juhul eksitakse pigem kontekstis, kus sarnaste sõnade juures on võimalik nii koma kasutus kui ka koma ärajätmine. Nii on näites (2) eksitud koma kasutamisega sidesõna *kui* ees: koma pannakse vaid siis, kui võrdluse teises pooles esineb verb.

(2) *Ei väike hiinlanna ei olegi parem**, kui** väike venelanna Kohtla-Järvelt.*

Eri keelte grammatikakorrektorites on valdavalt püütud lisada puuduolevaid komasid, vähematel juhtudel ka eemaldada üleliigseid. Nii Bick (taani keel) kui ka Izaskun Aldezabal jt (baski keel) pigem lisavad puudu olevaid kirjavahemärke (Bick 2006; Aldezabal jt 2003). Daniel Hardt (2001), kes püüdis masinõppemeetodil luua taani keele grammatikakorrektorit, tuvastas seevastu vaid üleliigseid komasid, kasutades õppe-

alusena ajalehetekste, kuhu oli suvaliselt komasid lisatud. Baski keele teisel, masinõppemeetodil loodud grammatikakorrektor tuvastab mõlemat tüüpi komavigu, kusjuures tulemused osutavad, et puuduolevate komade leidmine on märgatavalt keerulisem (Alegria jt 2006). Eesti keele puhul võtsin arvesse nii puuduolevaid kui ka üleliigseid komasid, kuid mõlemat tüüpi vaid teatud kontekstis. Teisisõnu kontrollisin komade esinemist seal, kus sagedamini eksitakse, ega arvestanud võimalike üleliigsete komadega muude sõnade ümbruses.

Üleliigsete komade korral piisab vea parandamiseks selle tuvastamisest: kui on teada, et koma on tekstis sobimatu, siis tuleb see vaid eemaldada. Ent kui mõni koma, mis tekstis olema peaks, on tegelikult ära jäetud, ei piisa parandamiseks teadmisest, et koma tuleb lausesse lisada, vaid peab ka teadma, kuhu see täpselt lisada. Mõnel juhul, näiteks koma nõudvate sidesõnade korral on seegi ülesanne küllaltki lihtsalt lahendatav. Teistel juhtudel võib osalausepiiri leidmine osutuda keerulisemaks. Näiteks on raske määrata, kus täpselt peaks kahe finiitse verbivormi vahel koma olema ja kas vahepealsed sõnad, iseäranis määrused, kuuluvad pigem esimesse või teise osalausesse. Sellest tulenevalt tegelen esialgu vaid komavigade tuvastamise ja vastavate märgendite lisamisega, parandus-te väljapakkumiseks läheb vaja täiendavat analüüsi. Samas on mõnel juhul küllaltki kerge juba vastavalt olemasolevale veamärgendile otsustada, kuidas lause õigekeelsust parandada.

Komavigade tuvastamisel on aluseks võetud grammatikakäsiraamatus (Erelt 2006) komakasutust käsitlevad õigekirjareeglid, tuvastaja otsib lausetest nii üleliigseid kui ka puuduolevaid komasid. Tuvastajasse on valitud need reeglid, mis määravad komakasutuse üheselt ära, s.t välja on jäetud olukorrad, kus on põhimõtteliselt õige nii komaga kui ka komata kasutus, millest ühte vaid hea stiili huvides soovitatakse. Nagu maini-

tud, ei ole sisse võetud kõik vastavad käsiraamatus esinevad reeglid, vaid üksnes need, mis normeerivad koma olemasolu või puudumist side- ja küsisõnade ning osalausete öeldiste vahel.

3. Lähenemine

Nagu eespool öeldud, kasutatakse grammatikakorrektorite juures ja keeletehnoloogias üldse nii reeglipõhiseid, statistilisi kui ka hübriidmeetodeid. Sealjuures on sama keele jaoks proovitud eri tüüpi lahendusi, kuigi enamasti mitte samade vealiikide puhul.

Eesti keele grammatikakorrektori loomisel on aluseks võetud reeglipõhine lähenemine ja kitsenduste grammatika formalism. Kitsenduste grammatika formalismi on välja töötanud Fred Karlsson Helsingi Ülikoolist, mõeldes sealjuures eelkõige süntaktilisele analüüsile. Hiljem on seda formalismi kasutatud ka paljudes muudes valdkondades, sealhulgas grammatikavigade märgendamisel. Reeglid tegelevad kas märgendite lisamise (lubades ühe sõna juures mitu eri märgendit) või õige märgendi määramisega, s.t sobiva valiku määramise või ebasobivate valikute eemaldamisega. Lisaks märgendatavale sõnale saab reeglis määrata kontekstimustri nii sõnade kui ka märgendite kujul, millal reeglit rakendada. Sealjuures on konteksti ulatuseks vaikimisi küll lause, kuid selle piire on võimalik vajadusel muuta. Kitsenduste grammatika reeglid on küllaltki arusaadavad ning hõlpsasti muudetavad, mis on üks reeglipõhise lähenemise eeliseid statistiliste või masinõppesüsteemide ees (Karlsson jt 1995: 42).

Üheks põhjuseks formalismi valikul oli asjaolu, et eesti keele automaatne morfoloogiline ning süntaktiline analüüs kasutavad sama meetodit, mis muudab eri tööriistade integreerimise

lihtsamaks. Et kitsenduste grammatikas on reeglid jaotatud eraldi moodulitesse, mida järjest rekursiivselt rakendatakse, siis on formalism iseäranis sobiv. See võimaldab muuhulgas vigade leidmise ja korrigeerimise jagamist eri moodulitesse või ka ühilduvus- ja komavigade järkjärgulist kontrolli, kasutades igal järgmisel analüüsiringil juba suurema kindlusega sooritatud parandusi. Kitsenduste grammatika puudusena võib välja tuua, et ehkki sõnade märgendeid on võimalik muuta ja sel moel näiteks ühilduvusvigu parandada, on formalism mõeldud eelkõige siiski analüüsiks ega võimalda sõnade (või kirjavahemärkide) lisamist ega eemaldamist. Selle lähenemise olen võtnud aluseks vigade leidmisel ja märgendamisel.

Kuna märgendada saab vaid lauses esinevaid sõnu, siis pole võimalik puuduolevale komale märgendit lisada. Seega on märgendamisele võetud pigem need sõnad, mille läheduses on oht komakasutuses eksida: sidesõnad ja küsisõnad ning samuti verbide pöördelised vormid, mille vahel peab alati leiduma kas sidesõna või koma. Märgendamise algul lisatakse kõigile kaks märgendit 'õige' ja 'väär', järgnevalt valitakse kitsendusreeglitega välja korrektne märgend. Siinkohal tähistab 'väär' olukorda, kus sõna ees esineb komaviga. Kuna lauses võib olla mitu märgendatavat sõna, võib osa neist saada ühe, teised teise märgendi, mis aitab selgitada vea asukohta. Arvestades, et eelistatum on olukord, kus grammatikakorrektoril jääb mõni keerulises lauses esinenud viga märkamata, kui liig sagedased valeteated, on reeglite koostamisel seatud eesmärgiks pigem võimalikult täpsed kui laiad reeglid ning võimalikult vähene valealarmide hulk.

See tähendab, et iga reegli puhul tuli vastavalt võimalikele eranditele täpsustada konteksti ning kaheldavas olukorras jätta alles pigem korrektset kasutust tähistav märgend. Nii

näiteks on ühendi *nii et* kasutuses lubatud koma panna nii ühendi ette kui ka vahele, kuid üldjuhul peab ühend lauses esinema koos komaga. Kuna aga erandjuhuna on lubatud püsiühendite (näiteks "*Valetab nii et suu suitseb.*") korral koma panemata jätta, siis tuli tuvastaja reeglites lubada ka komata kasutus, kuigi sel juhul võib osa puuduolevaid komasid leidmata jääda.

Reeglite koostamise aluseks olid grammatikaõpik, millest saab lause õigekeelsuse reegleid arvutiformalismi ümber kirjutada, ja tekstikorpused, millel reegleid testida ning mille põhjal katmata jäänud juhtumeid lisada. Kõik õigekirjareeglid kasutatud korpusel rakendust ei leidnud, kuid samas tuli välja mitu juhtumit, kus tegelik keelekasutus tingis vähemalt olemasolevate reeglite täpsustamise või neile erandjuhtumite lisamise, kuna automaatne analüüs vajas detailsemaid juhendeid.

4. Korpused

Reeglite koostamisel kasutasin kolme liiki tekstikorpusi, mis olid automaatselt morfoloogiliselt ja süntaktiliselt analüüsitud ning grammatikavead käsitsi märgendatud. Peamine korpus, mille põhjal uusi reegleid konstrueerida, koosneb Delfi internetiportaali foorumite kommentaaridest kogutud grammatiliselt vigastest lausetest, milles on kokku üle 9000 sõna. Tegu on nii teemade kui ka autorite poolest varieeruva tekstiga, mis võimaldab vaadelda erinevate inimeste keelekasutust ning eri eksimistasemeid õigekirjas. Nii on eespool toodud näites (1) kas tahtlikult või tahtmatult jäetud välja kõik komad, samal ajal kui näites (3) on autor osa komasid välja kirjutanud ning vaid kaks ära jätnud.

(3) *Praegu ongi nii **et** palu jumalat, et juhul kui satud süütult õnnetusse, siis süüdioleval osapoolele oleks norm kindlustaja (mitte Salva või Inges näiteks) – muidu jäädki uste vahet jooksma ja aega raiskama, õnnetuse pärast **milles** sina üldse süüdi pole.*

Kuna tegu on võrdlemisi spontaanse tekstiga, mida eelnevalt tõenäoliselt kuigivõrd ei kontrollita, siis on internetikommentaarides õigekirjavigade osakaal suhteliselt suur. Samas on tegu loomulike eksimustega, mitte konstrueeritud vigadega, nagu näiteks rootsi grammatikakorrektori treenimisel tehti (vt Hardt 2001). Grammatikakorrektori koostamisel on mõttekam kasutada võrdlusallikana tegelikku teksti, millega sarnasele hiljem tööriista rakendama hakatakse, kui tehislikult vigast teksti, mida korrektor hiljem tõenäoliselt ei kohta. Sarnastel põhjustel kasutati ka taani keelele grammatikakorrektori loomisel düsgraafikute kirjutatud suurte vigade osakaaluga tekste (vt Bick 2006).

Kuna üks oluline kriteerium reeglite koostamisel oli valealarmide vältimine, siis kontrollisin komavigade tuvastaja tööd ka ligi sajalauselisel Eesti keele koondkorpuse[2] osal. Eeldatavasti on publitseeritud tekstide õigekeelsus kontrollitud ning nii võis lihtsalt, vähese märgendamise vaevaga kasutada küllaltki suurt korpust. Kuigi uute reeglite koostamisel polnud kirjakeele korpusest suurt kasu, sai sellel kontrollida loodud reeglite töökindlust.

Võib arvata, et internetikommentaarid ei esinda just kõige tüüpilisemat kirjakeelt – laused on küllaltki pikad, sisaldavad palju osalauseid, ei pruugi olla struktureeritud samal moel kui tüüpiline tekstiredaktorisse kirjutatav tekst. Seetõttu kontrolliti komavigade tuvastaja reegleid ühel magistritööl (Liin 2008). Selgus, et teiste tekstiliikide põhjal koostatud reeglid toimivad

[2] http://www.cl.ut.ee/korpused/segakorpus/ (20.03.2009).

hästi ka tüüpilise arvutil kirjutatud teksti puhul. Programm leidis üles isegi ilmselt meediumipõhise tekkega vead nagu tekstiosa kopeerimise tõttu lausesse jäänud korduvad verbivormid, vt näide (4).

(4) *Muul juhul **peab** koma nõudva sidesõna ees **peab** olema koma.*

5. Tulemused

Komavigade tuvastaja koostamisel valmis 98 kitsendusreeglit, mida on küllaltki palju üksnes komavigade määramiseks. Sealjuures 4,5% juhtudest ei õnnestunud korrektset märgendit valida. Probleeme tekkis oodatult seal, kus komakasutus sõltub rohkemast kui morfoloogilisest infost ja arvesse tuleb võta ka kõrvallause sisu. Küllaltki sage on siinkohal selline juht, kus tuleb otsustada, kas osalause kuulub eelneva kõrvallause või pealause juurde – näites (5) tuleks *ja* ette koma panna just seetõttu, et sellele järgnev osalause on rinnastatud pealausega, viimast on aga automaatselt raske tuvastada.

*(5) Täna siis üks õnnetu opeli juht ei leidnud oma suunatule kangi üles kui reastus **ja** ühe kaubiku juhil puudus ka lisavarustuses suunatule kang.*

Teine suurem raskuste tekkepõhjus on asjaolu, et morfoloogilise analüüsi ja ühestamise tööriistad on mõeldud kasutamiseks keeleliselt korrektsetel lausetel ning seetõttu eksisid need vigaste lausete sõnadele analüüsil. Sääraseid analüüsivigu on võimalik mingil määral parandada, kui arvestada süntaktilise infoga, mis ongi üks grammatikakorrektori eesmärkidest. Lihtsaim juhtum on trükiviga, mis on tuvastatav kui keeles mitteesinev vorm; sõnade kokkukirjutamise või mõne muu sõnavormiga kattuvuse korral võib see aga probleemsemaks osutuda. Grammatikakorrektori segadusse ajamiseks piisab sellestki, kui pärast koma on tühik ära jäänud

(vt näide 6) – eelneva parandamiseta ei suuda märgendajad teist osalauset korrektselt märgendada ja eitussõna *pole* saab komavea märgendi.

(6) *Alguses öeldi kohe et väljaload unustage **ära,kui** just midagi pakilist **pole**.*

Teine vigase analüüsi põhjus on viga morfoloogilisel ühestamisel. Kuna morfoloogiline ühestaja arvestab otsuste tegemisel korrektse kontekstiga, siis võib see mõne ärajäänud või vale sõnavormi läheduses valida võimalikest analüüsidest tegelikult sobimatu. Nii on näites (7) analüüsitud sõnavormi *saate* kui nimisõna ilmselt just seetõttu, et sellele paistab samas osalauses eelnevat teinegi finiitne verbivorm.

(7) *Ise nad ju ütlesid et otsige süüa sealt **kust saate***

- *saate("saade+0" // _S_ com sg gen ADVL*

Lisaks võimalikule valele valikule võib morfoloogiline ühestaja põhjustada raskusi ka siis, kui liiga vähese info tõttu jäetakse sõnale alles mitmene analüüs, millest osa analüüse osutab lause ebakorrektsusele, teised aga sobivad olemasolevasse lausesse probleemideta. Näites (8) on lause analüüs raskendatud, kuna sõnal *täis* on alles jäetud ka verbi märgend.

(8) *Üks tüüp kes oli pidev hüppeskäia oli juba paar kuud teenistuse lõpetanud, kui juua **täis** peaga üle väeosa aia ronis.*

- "täis+0" // _A_ pos AN> PRD
- "täis+0" // _D_ ADVL
- "täis+0" // _S_ com sg nom SUBJ
- "täi+s" // _S_ com sg in ADVL NN>
- **"täi+s" // _V_ main indic impf ps3 sg ps af #FinV #InfP +FMV**

Määratud sõnade puhul oli komavigade tuvastaja täpsus 95% ja saagis 93%. Sellega oli täidetud üks algselt püstitatud eesmärk – viia ekslike veamärgendite hulk miinimumini. Lisaks õnnestus leida üllatavalt suur osa tekstides leidunud komavigadest. Lausetes tervikuna on tulemused veelgi paremad, kuna ühe sõna märgendamisel tehtud vea võib kompenseerida teise sõna märgend. Nii on näites (9) *vaid* märgitud ekslikult sidesõnaks, mille tõttu *mis* kohta käivad reeglid lubavad seal ekslikult komata kasutuse. Samas aga rakendub sidesõna *vaid* ette koma nõudev reegel ning lauses leitakse siiski komaviga. Samuti annaks alarmi sõna *on*, kuna reeglite kohaselt peab kahte pöördelist tegusõna eraldama kas koma mittenõudev sidesõna või koma.

(9) *Ta ju küsis **vaid mis** vahe **on** automaat- ja poolautomaatkäigukastiga autol?*

Nii õnnestuski vältida valealarme, kuigi korrektselt märkimata lauseid oli ligi 5%, mis tähendab, et 150 kontrollitud lausest, millest pooled sisaldasid süntaksivigu, jäi leidmata 4 vigast lauset.

Kokkuvõtteks

Komavigade tuvastamisel saavutatud täpsus näitab, et loodud tööriist on võrreldav muude keelte grammatikakorrektoritega: kitsenduste grammatikas kirjutatud Põhjamaade grammatikakorrektorite töötäpsused jäävad vahemikku 70–95% (vt Hagen jt 2001; Hagen jt 2002: 4). Samas tuleks arvestada, et teiste vealiikide lisamisel ning pärast vea tuvastamist sellele korrektse paranduse väljapakkumisel võib täpsus väheneda. Sellest hoolimata võib kindlalt väita, et antud lähenemine on eest keele grammatikakorrektori loomisel tulemuslikuks osutunud.

Nagu reeglite koostamisel selgus, oli korpuste põhjal uuritud tegelikest keelenäidetest grammatikakorrektori töö parandamisel palju kasu. Ka testimisel tuvastamata jäänud vigade hulgas oli selliseid, milles suurema hulga näidete kasutamisel oleks saanud eksimise välistada. Seega tuleks reeglite loomisel ja täpsustamisel edaspidi kasutada suuremat korpust ja kindlasti mitte piirduda üksnes üht tüüpi keelekasutusega, et vältida reeglite ülepassitamist ja sellest tulenevat sobimatust teistele keelekasutustele. Grammatikakorrektori loomisel on plaanis abiks võtta Tartu Ülikoolis kirjutatud bakalaureusetööde mustandversioone, milles kõik vead veel parandatud pole, ja keeleõppijate kirjutatud tekste, milles on samuti ohtrasti vigu.

Kui tekstis on komavead parandatud ja osalausepiirid paigas, siis saab grammatikakorrektorile lisada muude kirjavahemärkide kasutuse, ühildumise, rektsiooni, kokku-lahkukirjutamise, konteksti sobimatute sõnade (näiteks kirjavea tõttu vale tähenduse saanud sõnavormide) leidmise ja asendamise ning muudegi vealiikide määramise. Veatuvastusreeglite loomise käigus tuleb lisaks mõelda parandusettepanekute koostamisele, mida läheb vaja grammatikakorrektori automaatsel rakendamisel ning kasutaja töö kergendamiseks. Tekstiredaktoris tasub kirjutaja abivahendina mõelda stiilikorrektorile, mis aitaks vältida mitte üksnes otseseid vigu, vaid ka kordusi ning sobimatut sõnakasutust.

Kirjandus

Aldezabal jt 2003 = Aldezabal, Izaskun & Aranzabe Maxux & Arrieta Bertol & Maritxalar Montse & Oronoz Maite 2003. Toward a punctuation checker for Basque. ATALA workshop of punctuation (Paris). http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1069080468/publikoak/Toward-a-punctuation-checker-for-Basque.pdf (20.03.2009).

Alegria jt 2006 = Alegria, Iñaki & Arrieta, Bertol & Díaz de Ilarraza, Arantza & Izagirre, Eli & Maritxalar, Montse 2006. Using Machine Learning Techniques to Build a Comma Checker for Basque. Coling-ACL. Sydney. Australia, 1–8. http://ixa.si.ehu.es/Ixa/Argitalpenak/Artikuluak/1150185248/publikoak/komak-ML.pdf (20.03.2009).

Bick, Eckhard 2006. A Constraint Grammar Based Spellchecker for Danish with a Special Focus on Dyslexics. – SKY Journal of Linguistics. Vol. 19. http://www.ling.helsinki.fi/sky/julkaisut/SKY2006_1/1.6.1.%20BICK.pdf (20.03.2009).

Erelt, Mati 2006. Lause õigekeelsus. Juhatused ja harjutused. Bookmill.

Hagen jt 2001 = Hagen, Kristin & Lane, Pia 2001. "Det er fort gjort og skrive feil." En presentasjon av en automatisk grammatikkontroll for bokmål. Foredrag på Mons, Oslo. http://www.hf.uio.no/tekstlab/prosjekter/Mons.gr-sjekker.htm (20.03.2009).

Hagen jt 2002 = Hagen, Kristin & Johannessen, Janne Bondi & Lane, Pia 2002. The performance of a grammar checker with deviant language input [Proceedings of the 19th International Conference on Computational Linguistics.]. Vol. 2. Taipei, Taiwan. http://portal.acm.org/citation.cfm?id=1071884.1071894 (20.03.2009).

Hardt, Daniel 2001. Transformation-Based Learning of Danish Grammar Correction [Proceedings of RANLP 2001]. http://www.id.cbs.dk/~dh/papers/ranlp.pdf (20.03.2009).

Karlsson jt 1995 = Karlsson, Fred & Voutilainen, Atro & Heikkila, Juha & Anttila, Arto 1995. Constraint Grammar: A Language-independent System for Parsing. Berlin and New York: Walter de Gruyter. http://books.google.com/books?hl=en&lr=&id=70IvVPlH63cC&oi=fnd&pg=PP10&dq=Constraint+Grammar:+A+Language-independent+System+for+Parsing&ots=mA5pxnqGGb&-sig=ajX5aV_JUpdNp5FmhqpscP9UhIY (20.03.2009).

Liin, Krista 2008. Reeglipõhine komavigade tuvastaja eestikeelsetele tekstidele. Magistritöö. Juhendaja Kaili Müürisep. Tartu Ülikool, matemaatika-informaatikateaduskond, arvutiteaduse instituut. Tartu.

Grammar checker for detecting comma mistakes

Krista Liin

Summary

The aim of this grammar checker was to detect comma mistakes in written Estonian. As of yet, the checker does not suggest corrections, but that function could be added to the existing system later. The grammar checker rules are based on the Constraint Grammar Formalism framework.

The corpus used for rule development and testing consists of grammatically incorrect sentences gathered from the user postings on an Internet site. More than 9000 words were first morphologically and syntactically analyzed, and then manually tagged for comma error detection. Finite verb forms, interrogative words and conjugations were tagged and marked as correct or incorrect, depending on whether there was a comma mistake before those words.

The 98 constraint rules were tested on a 150-sentence test corpus of both incorrect and correct sentences. A precision of 95% and recall of 93% was achieved on tagged words. There were no sentence-level false alarms. The problems in detecting mistakes were mainly caused by incorrect spelling, previous tagging, or situations where the usage of comma depends on semantic information.

The results achieved are comparable to other grammar checkers. In the future, the grammar checker for Estonian will be further developed using larger corpora and targeting also other error types, such as agreement mistakes. The aim is to

test it also on the texts written by language learners and on other text types.

Autor

MSc Krista Liin, Tartu Ülikooli doktorant, Tartu Ülikooli arvutiteaduse instituudi projektijuht, krista.liin@ut.ee

KORPUSPOHJAINEN TUTKIMUS VIRONKIELISTEN SUOMEN-OPPIJOIDEN SISÄPAIKALLIS-SIJOJEN KÄYTÖSTÄ

Keaty Siivelt

Abstrakti

Tässä artikkelissa tarkastellaan vironkielisten suomenoppijoiden sisäpaikallissijojen käyttöä suomen oppijankielen korpuksessa (tästä lähtien ICLFI[1]). Artikkelissa analysoidaan, miten vironkieliset suomenoppijat käyttävät sisäpaikallissijoja ja millainen on mahdollinen lähdekielen vaikutus. Tutkimuksen lähtökohtana oli viron ja suomen sisäpaikallissijojen kontrastiivinen analyysi, jonka tuloksia on mahdollisen lähdekielen vaikutuksen selvittämiseksi verratu sisäpaikallissijojen korpuspohjaisen analyysin tuloksiin. Korpuspohjainen tutkimus takaa laajan aineiston vuoksi tulosten paremman yleistettävyyden. Tutkimuksen tuloksena selvisi, että vironkielisille oppijoille aiheuttavat eniten ongelmia suomen ja viron morfologian ja fonologian eikä morfosyntaksin väliset eroavaisuudet.

Avainsanat: oppijankieli, korpuspohjainen tutkimus, oppijankielen universaalit, kontrastiivinen analyysi, lähdekielen vaikutus, vironkieliset suomenoppijat

[1] ICLFI = International Corpus of Learner Finnish

1. Johdanto

Morfologisesti mutkikkaiden kielten välistä siirtovaikutusta on tutkittu vähän ja sen takia on vaikea tehdä morfologisesta siirtovaikutuksesta minkäänlaisia johtopäätöksiä. Hyvän mahdollisuuden siihen antaa kahden morfologisesti mutkikkaan ja läheisen sukukielen suomen ja viron oppijankielten tutkiminen, koska läheisissä sukukielissä on kieliopillisesti laajempi kosketuspinta ja sen myötä myös suurempi mahdollisuus siirtovaikutukseen (Kaivapalu 2008: 104). Nimenomaan läheisen sukukielen oppimisessa on merkittävä tekijä juuri kielten välinen yhtäläisyys (Ringbom 2007: 1). Oppijankielen tutkiminen antaa mahollisuuden tutkia sekä lähdekielen vaikutusta että myös oppijankielelle itselleen ominaisia piirteitä. Oppijankielen tutkimus on ensinnäkin merkittävä syystä, että sen avulla voi muokata kielen opettamisprosessia kielen oppijan tarpeisiin (Eslon 2007: 98).

Tässä työssä tutkitaan lähdekielen vaikutusta toisen kielen oppimiseen. Kielten välisten eroavaisuuksien ja samanlaisuuksien selvittämiseksi pohjauduin Eve Kaskin diplomityöhön ”Suomen ja viron sisäpaikallissijojen kontrastiivinen analyysi“ (1991). Kielten kontrastiivisen analyysin tuloksista ilmeni, että kielten sisäpaikallisisjojen käytössä on sekä eroavaisuuksia että yhtäläisyyksiä, joka antaa mahdollisuuden tutkia sekä myönteistä että kielteistä lähdekielen vaikutusta. Kontrastiivisen analyysin tulokset eivät yksinään anna varmoja tuloksia siitä, millaisia ovat ne kohdat kielen käytössä, jotka aiheuttavat oppijoille ongelmia. Todellisen kuvan oppijoiden kielenkäytöstä saa heidän autenttisten tuotosten analysoinnista. Korpusaineisto antaa siihen hyvän mahdollisuuden ja myös laajan pohjan. Artikkelissa tarkastellaan miten kontrastiivinen analyysi ja oppijankielen korpuspohjainen analyysi johtavat lähdekielen vaikutuksen selvittämiseen vironkielisten suomen-

oppijoiden sisäpaikallissijojen käytössä. Tuloksista selvisi että lähdekielen vaikutuksen kannalta ovat olennaisia kaikkien kielitasojen väliset samanlaisuudet ja eroavaisuudet.

2. Kontrastiivinen analyysi, lähdekielen vaikutus ja toisen kielen oppiminen

Kontrastiivinen analyysi oli ensimmäinen yritys ratkaista toisen tai vieraan kielen oppimisen kysymyksiä. *Toinen kieli* merkitsee kieliympäristössä opittavaa kieltä ja *vieras kieli* kieliympäristön ulkopuolella opittavaa kieltä. Perinteisessä kontrastiivisessa analyysissa verrataan yleensä kahden kielen rakenteita kielten erilaisuuksien ja samankaltaisuuksien paljastamiseksi. Nykyaikainen soveltava kontrastiivinen tutkimus pyrkii tehostamaan kielenopetusta (Sajavaara 1999: 106–110). Toisen kielen oppimisen kannalta kontrastiivisella analyysilla on merkittävä tarkoitus. Nimittäin strukturalistis-kontrastiivisen teorian vahvan hypoteesin mukaan on kontrastiivisen analyysin pohjalta mahdollista ennustaa oppijan vaikeuksia kohdekielen oppimisessa. Tässä kohtaa täytyy silti mainita että ongelmien aiheuttajina kohdekielessä on myös muita tekijäitä kuin kohdekielen rakenteiden poikkeaminen lähdekielestä. Kohdekielen käyttöön vaikuttavat mm. syötös (*input*), kielen opetus ja monet individuaaliset tekijät kuten ikä, motivaatio, kielikyky, aiemmin opitut kielet jne (Larsen-Freeman & Long 1991: 204–206). Kontrastiivinen analyysi tuo esiin sekä kielten väliset eroavaisuudet että samanlaisuudet, jotka luovat hyvän mahdollisuuden siirtovaikutukselle. Kun lähde- ja kohdekielet konvergoivat, tuloksena voi olla positiivinen siirtovaikutus tai negatiivinen divergenssi kun lähde- ja kohdekielet divergoivat. Siirtovaikutus itsestään ei voi olla positiivista eikä negatiivista vaan sen lopputulosta voi pitää kohdekielen kannalta positiivisena tai negatiivisena (Sajavaara 2006: 11).

Siirtovaikutuksesta käytetään yleisimmin englanninkielistä termiä *transfer*. Olennaista on silti huomauttaa että transferin nimitystä on ennenkaikkea kytketty behavioristiseen näkökulmaan. Kielen oppimisen pääongelmana nähtiin äidinkielen rakenteiden ilmenemistä kohdekielen rakenteiden sijasta. Lähdekielen vaikutusta toisen tai vieraan keielen oppimiseen alettiin kuvata nimellä *transfer*. Kohdekielen kannalta negatiivista siirtovaikutusta alettiin kutsua *interferenssiksi*. Perusluonteeltaan on sekä positiivisen että negatiivisen lähdekielen vaikutuksen kannalta kyse samasta prosessista: yhdestä kielestä oletetaan siirtyvän tai siirrettävän jotakin toiseen (Kaivapalu 2006: 30).

Tässä artikkelissa *lähdekielen vaikutukseksena* nähdään Scott Jarvisin mukaan sellaiset oppijakielen käyttötapaukset, joissa oppijoiden kohdekielisen käyttäytymisen jonkin piirteen ja heidän lähdekielisen taustansa välillä voidaan osoittaa olevan tilastollisesti merkitsevä tai todennäköisyyteen perustuva (*probability based relation*) korrelaatio (Jarvis 2000: 252).

Lähdekielen vaikutus on toisen kielen omaksumisen ja vieraan kielen oppimisen tutkimuksessa eniten käsitelty aihe. Suurin osa lähdekielen vaikutuksen tutkimuksista on kohdistunut indoeurooppalaisiin kieliin ja sen myötä on myös kehittynyt kuva miten ja milloin lähdekieli vaikuttaa kohdekielen oppimiseen (Kaivapalu 2008: 94). Morfologisesti mutkikkaiden kielten välistä vaikutusta on tutkittu vähän, osittain siitäkin syystä on jopa väitetty ettei siirtovaikutusta esimerkiksi morfologiassa tapahdukaan (Jarvis & Odlin 2000: 536). Lähdekielen vaikutusta on enimmäkseen pidetty myös negatiivisena ilmiönä, siitä arvellaan johtuvan kohdekielen virheet. Vähemmän on puhuttu lähdekielen myönteisestä vaikutuksesta. Vasta viime vuosina on alettu puhua kieltenvälisen samanlaisuuden (*cross-linguistic similarity*) tärkeydestä (Ringbom 2007: 1–2).

Jos kohdekieli ja oppijan äidinkieli ovat läheisiä sukulaiskieliä, niin tämä voi juuri opintojen alkuvaiheessa auttaa kielenoppijaa (Ringbom 2007: 1), sillä oppija pystyy hahmottamaan sanoja ja rakenteiden tehtäviä lähdekielen avulla. Joskus tämä aiheuttaa ongelmia, esimerkiksi kohdekielessä vältetään ilmauksia, jotka ovat lähde- ja kohdekielessä samanlaisia, vain siitä syystä, että ne tuntuvat liian tutuilta (Sajavaara 2006: 21) (Kaivapalu 2005: 250).

Lähdekielen vaikutus on vain yksi toisen kielen omaksumiseen vaikuttavista tekijöistä. Kehitys sai alkunsa ns. Krashenin monitorimallista, jonka mukaan ihmisellä on monitori, joka tarkistaa sekä tietoista että tiedostamatonta kielenkäyttöä (Kaivapalu 2006). Krashenin mukana opitaan syötöksestä (*input*) ja kaikki kielen tuottaminen alkaa omaksutulta pohjalta. Olennaista on syötös. Oppimisen esimerkiksi luokkatilanteessa arveltiin olevan syötöksen vahvistamista ja lisäävän sen myötä opittavien asioiden selkeyttä (Sajavaara 2006: 20). Silti epäselväksi jäi miten oppijat hahmottavat opittavan kielen elementtejä. Yksi mahdollisista selityksistä on lähdekielen vaikutus. Kontrastiivinen analyysi voi toimia syötöstä vahvistavana, tuodessaan esiin lähdekielen ja kohdekielen väliset eroavaisuuet (Sajavaara 2006: 22), mutta myös tukemalla lähdekielen malleja, jotka johtavat kohdekielessäkin hyväksyttävään tulokseen.

3. Oppijankielen korpuspohjainen analyysi

Termit *oppijankieli* ja *välikieli* merkitsevät molemmat kielen varianttia jonka kielen oppijat luovat kohdekielessä. Termin välikieli (*interlanguage*) otti käyttöön L. Selinker vuonna 1972 ja sitä on kytketty sekä behavioristiseen kielen käsittelyyn, kontrastiiviseen tutkimukseen että interferenssiteoriaan. Termi

oppijankieli (*learner language*) on käytettävissä toisen ja vieraan kielen omaksumisen (*second/foreign language acqusition*) viitekehyksessä (Eslon 2007: 87–88).

Korpus (*corpus*) on suunitelmallisesti koottu, laaja sähköinen aineisto, jota on mahdollista hyödyntää sekä kielen tutkimisessa että kielen opetuksessa (Jantunen 2008). Oppijankielen korpus koostuu teksteistä, joita oppijat luovat kirjallisesti kohdekielessä. S. Grangerin (2003) mukaan oppijankielen korpuksen (*learner corpus)* lisäksi ovat käytössä sekä termi välikielenkorpus (*interlanguage corpus*) että toisen kielen korpus (*L2 corpus*). Kaikenlaista korpusaineistoa on mahdollista analysoida sekä manuaalisesti, että erillaisten tietokoneohjelmien avulla (esim. Oxford WordSmith Tools). Korpusaineisto tarjoaa mahdollisuuksia sekä kvalitatiiviseen että kvantitatiiviseen aineiston käsittelyyn.

Oppijankielen kontrastiivisessa analyysissa (*contrastive interlanguage analysis – CIA*) vertaillaan kvalitatiivisesti ja kvantitatiivisesti oppijankieltä kohde- ja lähdekieleen ja/tai erillaisiin oppijankielen variantteihin (Metslang 2007: 141). Oppijankielen yksi mahdollisista kontrastiivisista tutkimustyypeistä on oppijankielen analyysi integroituna perinteiseen kontrastiiviseen tutkimukseen. Kontrastiivisen analyysin (*contrastive analysis – CA*) tuloksien tarkastelu korpuspohjaisesti antaa enemmän mahdollisuuksia poikkeuksien selittämiseen: onko kyseessä lähdekielen ja kohdekielen välinen siirtovaikutus, jokin muu universaali oppijankielen piirre tai jotain sangen muuta.

4. Aineisto ja tutkimusmenetelmä

Artikkelin aineisto on kerätty ICLFI:stä. Korpusaineiston olen valinnut juuri siitä syystä, että se laajan aineiston vuoksi takaa tulosten paremman yleistettävyyden ja sen analysointi on

tulevaisuudessa mahdollista myös erillaisilla tietokoneohjelmilla. Koska korpusaineiston keruu on vasta alkuvaiheessa, on kyseessä pilottitutkimus.

ICLFI on Oulun yliopistossa koottava oppijankielen korpus. Korpuksen aineisto koostuu suomen kieltä vieraana kielenä oppivien kirjallisista teksteistä. Korpusaineistoon kuuluu tällä hetkellä 17 erilaista lähdekieltä puhuvien oppijoiden tekstejä. Tekstien lisäksi korpukseen on koottu taustatietoja oppijoista (sukupuoli, äidinkieli, kohdekielen osaamisen taso, ikä jne.), jotka ovat olennaisia taustatekijöitä lähdekielen vaikutuksen tutkimisessa (Jarvis 2000: 245–261).

Korpustekstit on luokiteltu tasoittain (alkeistaso, keskitaso, edistyneet) ja tekstilajittain (hakemus, sarjakuva, kortti, vastine, uutinen, kirje, mielipide, referaatti, arvostelu, kertomus, päiväkirja, essee, kuvaus). Opiskelijoiden taso on määritelty opiskeltuun tuntimäärään perustuvaa kriteeriä käyttäen. Alkeistason opiskelijoiden opetukseen kulunut tuntimäärä on alle 200, keskitaso 200–400 tuntia ja yli 400 tuntia opetusta saaneet luokitellaan edistyneen tason oppijoiksi (Jantunen 2008b).

Artikkelissa keskityn juuri viroa äidinkielenään puhuvien suomenoppijoiden sisäpaikallissijojen käyttöön. Vironkielisten suomenoppijoiden tuottamia saneita on korpuksessa yhteensä 36.636.

Lähdekielen vaikutuksen tutkimiseksi pohjauduin ”Viron ja suomen sisäpaikallisijojen kontrastiiviseen analyysiin“ (Kask 1991). Kask vertasi kontrastiivisessa analyysissa suomen ja viron sisäpaikallissijojen käyttöä funktioittain. Selvitin sisäpaikallisijojen funktioiden väliset vastaavuudet suomessa ja virossa saadakseni kuvan oppijoille mahdollisesti ongelmia aiheuttavista funktioiden välisistä eroavaisuuksista ja samanlaisuuksista. Oppijankielen korpusaineiston pohjalta selvitin

miten vironkieliset suomenoppijat käyttävät sisäpaikallissijoja ja millaisia norminvastaisia poikkeuksia heidän sisäpaikallissijojen käytössään esiintyy. Lähdekielen vaikutusta on melkein mahdotonta nähdä virheettömästä tuotoksesta, kun taas virhe saattaa paljastaa lähteensä. Oppijan virheet ovat tärkeää tutkimisaineistoa, erityisesti kielten välisen vaikutuksen selvittämisessä niillä on merkitystä (Martin 1995: 264).

Lähdekielen vaikutuksen tutkiminen oppijankielen korpusten pohjalta on olennaista, koska lähdekielen vaikutus kuuluu oppijankielen universaaleihin piirteisiin (Kaivapalu 2008: 103).

5. Sisäpaikallissijojen käyttö oppijankielen korpuksen pohjalta

Tässä luvussa tarkastelen vironkielisten suomenoppijoiden sisäpaikallisijojen käyttöä suomen oppijankielen korpuksessa.

a. Illatiivi

Eve Kaskin tutkimuksessa ”Viron ja suomen sisäpaikallisijojen kontrastiivinen analyysi“ (1991) on esitelty 26 mahdollista illatiivin funktiota suomessa.

Illatiivia esiintyi vironkielisten suomenoppijoiden teksteissä 12 funktiossa, niistä 8 funktiossa jotka ovat suomessa ja virossa samanlaisia ja 4 sellaisissa jotka ovat suomessa ja virossa erilaisia (taulukko 1). Illatiivia verbirektion sijana voi virossa vastata sekä illatiivi että jokin muu rakenne. Funktioiden varsin pieni käyttömäärä on oppijankielelle ominaista.

Taulukko 1. Illatiivin vironkielisten suomenoppijoiden teksteissä

Samanlaiset funktiot suomessa ja virossa (n=415)	**Erilaiset funktiot suomessa ja virossa (n=78)**
1) Paikka, johon saavutaan tai mennään (271) 2) Paikka, johon mahtuu jotain (12) 3) Toiminta, työ 3. infinitiivi (96) 4) Paikka, johon joku haluaa (6) 5) Olotila, johon jotain laitetaan (2) 6) Paikka, johon jää jotain (2) 7) Suunta (5) 8) Illatiivi verbirektion sijana (kohde, aihe, syy) (15) 9) Norminvastainen käyttö (6)	1) Raja, johon jotain ulottuu (1) 2) Tilanne, johon mennään (49) 3) Aikakauden loppu, illatiivi + saakka, asti (1) 4) Illatiivi verbirektion sijana (kohde, aihe, syy) (18) 5) Norminvastainen käyttö (10)

Illatiivin käyttö enimmäkseen isossa määrässä funktioissa, jotka ovat suomessa ja virossa samanlaisia ja myös pieni määrä norminvastaisia käyttötapauksia samoissa funktioissa, osoittaa positiivista lähdekielen vaikutusta. Oppijan on todennäköisesti helpompaa käyttää illatiivia niissä funktioissa, jotka ovat lähde- ja kohdekielessä samanlaisia. Kyseisestä aineistosta ilmenee hyvin, miten kahdeksasta suomessa ja virossa samanlaisesta funktiosta neljä ilmaisevat yleisesti paikkaa ja muodostavat koko aineiston illatiivisijaisista saneista yli puolet. Kyseessä voi silti olla myös suomen kielelle ominainen piirre, jonka mukaan myös suomea äidinkielenään puhuvat käyttävät illatiivia yli puolessa käyttötapauksista paikan ilmaisuissa (Sajavaara 1999: 116).

Toiseksi eniten esiintyy suomenoppijoilla illatiivia toiminnan ja työn ilmauksissa (kolmas infinitiivi), jota vastaa virossa ma-infinitiivi. Suomen kolmas infinitiivi ja viron ma-infinitiivi ovat käytössä samoissa funktioissa (Siivelt 2008: 12–13). Esimerkiks *kun se rupeaa liikku**maan*** ja *kui see liiku**ma** hakkab*.

Kolmanneksi eniten esiintyi illatiivi vironkielisillä suomenoppijoilla tilanteeseen, tapahtumiin menemistä tai jonkun luonna olemista ilmaisevissa funktioissa. Esimerkiksi *he menivät hautajai**siin*** ja *nad läksin matuste**le***. Viron kieliopissa vastaa suomen illatiivin käyttöä kyseisessä funktiossa allatiivi. Todennäköisesti on oppijoilla kyseinen funktio helposti muistettavissa, koska kyseessä on suomen kielessä hyvin produktiivinen ja laajasti käytettävissä oleva funktio (Kask 1991:11).

Norminvastaisia käyttötapauksia eiintyi hieman enemmän funktioissa jotka ovat suomessa ja virossa erilaisia.

b. Inessiivi

Eve Kask (1991) esitti työssään 19 mahdollista inessiivin funktiota. Inessiivi esiintyi 4 oppijan teksteissä neljässä funktiossa. Vironkielisillä suomenoppijoilla ei esiintynyt sellaisia inessiivin funkioita, jotka olisivat olleet erilaisia virossa (taulukko 2).

Taulukko 2. Inessiivi vironkielisten suomenoppijoiden teksteissä

Samanlaiset funktiot suomessa ja virossa (n=1729)
1) Olemisen, toiminnan ja tapahtumapaikan merkityksessä (1576)
2) Työ, toiminta (3. infinitiivi) (17)
3) Aika, jolloin jotain tapahtuu (42)
4) Tilan ilmauksissa (23)
5) Norminvastainen käyttö (71)

Inessiivi oli tutkimassani aineistossa eniten käytetty sisäpaikallissija. ISK:n mukaan inessiivi on myös suomea äidinkielenään puhuvien teksteissä eniten käytetty sisäpaikallisija (ISK§1227). Kyseessä voi siis olla joko suomen kielelle ominainen piirre tai inessiivisijan yliedustus joka on oppijankielen universaali piirre (Jantunen 2008a: 70). Kumpikaan selitys ei ole toista pois-sulkeava. Kohdekielessä eniten käyttössä olevat rakenteet jäävät toistamisen myötä paremmin oppijan muistiin ja siitä johtuen voi juuri niiden käyttöfrekvenssi lisääntyä. Mitä opitaan enemmän, sitä todennäköisesti myös tuotetaan enemmän. Inessiivisijan käytössä ovat vironkieliset oppijat tuottaneet myös eniten norminvastaisia muotoja.

c. Elatiivi

Elatiivin funktioita oli Kask (1991) esittänyt yhteensä 16. Elatiivi esiintyi oppijoiden teksteissä 11 erilaisessa funktiossa, joista 10 funktiota vastasi myös virossa elatiivi ja yhtä voi virossa vastata sekä elatiivi että jokin muu rakenne (taulukko 3).

Taulukko 3. Elatiivi vironkielisten suomenoppijoiden teksteissä

Samanlaiset funktiot suomessa ja virossa (n=401)	**Erilaiset funktiot suomessa ja virossa (n=256)**
1) Paikka (239) 2) Aines ja alkuperä (9) 3) Syy (12) 4) Kokonaisuus, jonka osasta on kyse (12) 5) Vertailu (4) 6) Mielipidettä ja suhtautumista ilmaisevissa funktioissa (43) 7) Aika (12) 8) Kosketuskohta (1)	1) Elatiivi objektin sijana (rektiosijana) (231) 2) Norminvastainen käyttö (25)

9) Tuloslauseen lähtökohta (11) 10) Aihe (38) 11) Elatiivi objektin sijana (rektiosijana) (13) 12) Norminvastainen käyttö (8)	

Suomessa mahdollisista sisäpaikallissijojen funktioista ovat vironkieliset suomenoppijat käyttäneet eniten elatiivin funktioita. Syynä siihen on todennäköisesti suomen ja viron elatiivin funktioiden vastaavuus. 16 mahdollisesta funktiosta 10 vastaa myös virossa elatiivi.

Elatiivin erilaisten funktioiden käyttöön vaikuttaa todennäköisesti myös sijamuotojen samankaltaisuus, eli samanlaisuus morfologian tasolla. Jälkitavun pitkän vokaalin puuttuminen elatiivissa muuttaa myös sijan fonologian tasolla viroa lähemmäksi. Elatiivin funktioiden laajempi variaatio johtuu siis kielen eri tasoilla olevista yhtäläisyyksistä, joiden liittymisen tuloksena oppijalla vahvistuu tieto elatiivin käyttämisestä. Suomen ja viron elatiivin sijapäätteet ovat -stA ja -st. Kilpailevat muodot kummassakin kielessä puuttuvat. Kuten sanottu, niin myös fonologia on kyseisessä sijassa samanlaista, eli puuttuu virolaisille vaikea jälkitavun pitkä vokaali. Ongelmia voi aiheuttaa vain suomen vokaaliharmonia. Elatiivisijan funktiot kielissä ovat myös enimmäkseen samanlaisia. Kyseessä on siis todennäköisesti lähdekielen positiivinen vaikutus.

Vironkieliset oppijat ovat käyttäneet elatiivia eniten suomessa ja virossa erilaisissa funktioissa objektin sijana (verbirektion sijana), verbin *pitää* yhteydessä, jota virossa vastaa verbi *meeldima* + nominatiivi. 244 saneesta oli *pitää* verbistä johtuen elatiivisijaisia 190 sanetta. Loput saneista jakautuivat verbien *tykätä, kiinnostua, riippua* ja *unelmoida* välillä. Eniten norminvastaisia elatiivimuotoja johtui myös *pitää* verbin rektiosta.

Siitä huolimatta näkyy, että *pitää* verbin rektiosijana on vironkielisillä suomenoppijoilla vakiintuneena elatiivisija.

d. Sisäpaikallissijojen norminvastaiset käyttötapaukset

Norminvastaisiksi olen luokitellut kaikki ne tapaukset jotka eivät vastaa suomen yleiskielen normeja. Norminvastaiset käyttötapaukset ovat luokiteltu kielitasoittain ja tyypeittäin. Norminvastaiset tyyppit ja esimerkkilauseet ovat esitelty liitteessä 1.

Yhteensä lyötyi teksteistä 19 mahdollista norminvastaista tyyppiä. Niistä 8 luokittelin morfosyntaksin tason tapauksiksi: 5 tyyppiä samoissa funktioissa, 3 erilaisissa funktioissa; 8 morfologian tason tapauksiksi ja 3 fonologian tason tapauksiksi.

Illatiivin käytössä esiintyi eniten norminvastaisia tapauksia morfosyntaksin tasolla. Oppijoiden teksteissä esiintyvä norminvastainen inessiivimuodon käyttö illatiivimuodon asemasta funktioissa, joissa illatiivi ilmaisee paikkaa johon mennään tai saavutaan ei todellisuudessa selity suomen ja viron funktioeroilla. Paikan ilmaiseminen suomessa ja virossa on illatiivissa samanlaista. Todennäköisesti on kyseessä silti kielten morfologian tasolla olevasta erosta, jonka tuloksena on lähdekielen negatiivinen vaikutus kohdekielen morfosyntaksin tasolla. Suomen illatiivipäätteen asemasta käytetään viron illatiivipäätettä *-sse* muistuttavaa suomen inessiivipäätettä *-ssA.* Joissakin tapauksissa käytetään myös viron kieliopista peräisin olevaa päätettä *-sse.* Esimerkiksi *sitten pukeun, suutelen tyttoani ja lähden yliopisto**ssa*** tai *kun luennot ovat ohi, menen kahvila**sse** syömään.* Paradigmojen sekoittumisen tuloksena on lähdekielen negatiivinen vaikutus morfologian tasolla paikoissa, joissa se olisi morfosyntaksin tasolla positiivista.

Paradigmojen sekoittumisen syynä on myös sijapäätteen *-s* käyttö inessiivipäätteen *-ssA* asemasta, esimerkiksi *yliopisto**s** olen kaksitoistaksa neljaan.* Norminvastaiset muodot paikan ilmauksissa johtuvat suomen ja viron morfologisista eroista, jonka tuloksena on negatiivinen vaikutus morfosyntaksin tasolla. Toiseksi eniten ongelmia aiheuttivat suomen ja viron paikallissijojen distribuutioerot. Allatiivin käyttö inessiivisijan asemasta paikkaa ilmaisevissa funktioissa ja inessiivin käyttö allatiivin asemasta johtuvat suomen ja viron sisä- ja ulkopaikallissijojen jakaumaeroista. Suurimman osan norminvastaisista muodoista tuotetaan paikannimistä, joiden välillä ei ole tiukkaa sisä- ja ulkopaikallissijojen jakaumaa. Toiseksi eniten ongelmia aiheuttavat sanat *tässä* ja *täällä*. Tämä johtuu suomen ja viron paikallissijojen merkityseroista. Virossa nimittäin puuttuu paikan ilmauksissa sisäpaikallissijoilla täsmentävä merkitys (Kask 1991: 153).

Suomen ja viron taivutusjärjestelmien sekoittumisen syynä voi nähdä myös sijapäätteen *-ma* käyttöä 3. infinitiivin illatiivipäätteen *-mAAn* asemasta. Esim. *Nukku**ma** menen oikein myöhään, välillä puoli kaksi*. Todennäköisesti päätteiden samankaltaisuus ja eroavaisuus fonologian tasolla johtaa kyseiseen norminvastaiseen tuotokseen. Tässä kohtaa on kyseessä morfologian tasolla oleva poikkeus yleiskielen normeista, mikä johtuu kielten fonologisista eroavaisuuksista. Mahdollinen positiivinen lähdekielen vaikutus morfosyntaksin tasolla muuttuu kielteiseksi vaikutukseksi morfologian tasolla fonologian eroista johtuen. Fonologian eroista johtuvat myös poikkeamat illatiivimuodoissa, joissa puuttuu vokaali sanoissa, joiden illatiivi muodostetaan sijapäätteellä vokaali + *-n*. Sijapäätteen vokaali + *-n* käyttö illatiivipäätteen *-seen* asemasta johtuu todennäköisesti taivutustyyppien sekoittumisesta oppijan päässä. Esim. *Ensiksi menen vessaan, sitten kylpyhuonee**n**.*

Vironkieliset suomenoppijat käyttävät enemmän ja saavat myös enemmän positiivisia tuloksia niissä sisäpaikallissijojen funktioissa, jotka ovat suomessa ja virossa samanlaisia. Morfosyntaksin tasolla on havaitavissa myönteinen lähdekielen vaikutus. Yleiskielen normeista poikkeavat muodot johtuvat todennäköisesti lähdekielen negatiivisesta vaikutuksesta morfologian ja fonologian tasoilla, tukahtuttaen lähdekielen positiivisen vaikutuksen morfosyntaksin tasolla.

6. Päätelmiä

Artikkelin tavoitteena oli kartoittaa vironkielisten suomenoppijoiden sisäpaikallissjojen käyttöä ICLFI:n pohjalta ja tarkastella, miten lähdekieli mahdollisesti vaikuttaa siihen. Tutkimuksesta selvisi, että vironkieliset suomenoppijat käyttävät sisäpaikallisisjoja eniten funktioissa, jotka ovat suomessa ja virossa samanlaisia. Oppijat käyttivät sisäpaikallissijoja vähemmissä funktioissa, kuin suomen kielessä on mahdollista käyttää. Norminvastaisia tapauksia esiintyi myös sellaisissa funktioissa, jotka ovat suomessa ja virossa samanlaisia.

Artikkelista selvisi, että vironkielisten suomenoppijoiden teksteissä aiheuttavat eniten norminvastaisia muotoja kielten morfologian ja fonologian eroavaisuudet, eikä niinkään morfosyntaksin eroavaisuudet. Ongelmia aiheuttivat myös sisä- ja ulkopaikallissijojen distribuutioerot.

Lopuksi voi sanoa, että kielitasot ovat keskenään tiiviisti sidoksissa. Samanlaisuudet kielten eri kielitasoilla vahvistavat mahdollista positiivista lähdekielen vaikutusta.

Lähteet

Eslon, Pille 2007. Õppijakeelekorpused ja keeleõpe. – Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine. Eesti filoloogia osakonna toimetised 9 / Toim. P. Eslon. Tallinn: TLÜ Kirjastus, 87–120.

ISK = Hakulinen, Auli & Vilkuna, Maria & Korhonen, Riitta & Koivisto, Vesa & Heinonen, Tarja Riitta & Alho, Irja 2004. Iso suomen kielioppi. Helsinki: Suomalaisen Kirjallisuuden Seura. http://scripta.kotus.fi/visk (9.10.2008).

Jantunen, Jarmo 2008a. Haasteita oppijankielen korpusanalyysille: oppijankielen universaalit. – Õppijakeeleanalüüs: võimalused, probleemid, vajadused. Eesti filoloogia osakonna toimetised 10 / Toim. P. Eslon. Tallinn: TLÜ Kirjastus, 67–92.

Jantunen, Jarmo 2008b. Suomen oppijankielen korpus [luentomateriaalit]. Käsikirjoitus kirjoittajan hallussa.

Jarvis, Scott & Odlin, Terence 2000. Morphological type, spatial reference, and language transfer. – Studies on Second Language Acqusition 22, 535–556.

Jarvis, Scott 2000. Methodological Rigor in the Study of Transfer: Identifying L1 Influence in the Interlanguage Lexicon. – Language learning 50 (2), 245–349.

Kaivapalu, Annekatrin 2005. Lähdekieli kielenoppimisen apuna. Jyväskylä Studies in Humanities 44. Jyväskylä: Jyväskylän Yliopisto. dissertations.jyu.fi/studhum/9513923916.pdf (08.10.2008).

Kaivapalu, Annekatrin 2006. Suomi toisena/vieraana kielenä. [luentomateriaalit]. Käsikirjoitus kirjoittajan hallussa.

Kaivapalu, Annekatrin 2008. Lähtekeele mõju korpuspõhine uurimine. – Õppijakeeleanalüüs: võimalused, probleemid, vajadused. Eesti filoloogia osakonna toimetised 10 / Toim. P. Eslon. Tallinn: TLÜ Kirjastus, 93–19.

Kask, Eve 1991. Soome ja eesti kohakäänete kontrastiivne analüüs. Diplomitöö. Tarton yliopisto. Suomalais-ugrilaisten kielten laitos.

Larse-Freeman, Diane & Long, Michael H. 1991. An Introduction to Second Language Acquisition Research. New York: Longman.

Martin, Maisa 1995. The Map and the Rope.Finnish Nominal Inflection as a Learning Target. Studia Philologica Jyväskyläensia 38. Jyväskylä: University of Jyväskylä.

Metslang, Helena 2007. Õppijakeele korpuspõhisest analüüsist – Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine. Eesti filoloogia osakonna toimetised 9 / Toim. P. Eslon. Tallinn: TLÜ Kirjastus, 139–151.

Ringbom, Håkan 2007. Cross-linguistic Similarity in Foreign Language Learning. Clevedon: Multilingual Matters LTD.

Sajavaara, Kari 1999. Kontrastiivinen kielentutkimus ja virheanalyysi.- Kielenoppimisen kysymyksiä / Toim. K. Sajavaara &ja A. Piirainen-Marsh. Soveltavan kielentutkimuksen keskus. Jyväskylä: Jyväskylän yliopisto, 103–128.

Sajavaara, Kari 2006. Kontrastiivinen analyysi, transferi ja toisen kielen oppiminen. – Lähivertailuja 17. Jyväskylä Studies in Humanities 53 / Toim. A. Kaivapalu & K. Pruuli. Jyväskylä: Jyväskylän yliopisto, 9–25.

Siivelt, Keaty 2008. Suomen ja viron sisäpaikallissijojen kontrastiivinen analyysi oppijansuomen tutkimuksen pohjana. Seminarityö. Tallinnan yliopisto. Viron kielen ja kulttuurin instituutti.

Liite 1.

Taulukko 4. Norminvastaiset sisäpaikallisisjojen käyttötapaukset vironkielisten suomenoppijoiden teksteissä (MS = morfosyntaksin tason poikkeus, M = morfologian tason poikkeus, F = fonologian tason poikkeus, s = samanlaiset funktiot suomessa ja virossa, e = erilaiset funktiot suomessa ja virossa)

Sisäpaikallissijojen norminvastaiset käyttötapaukset	
Käyttötapaus	**Esimerkki**
1) Inessiivimuodon käyttö illatiivimuodon asemasta funktioissa, joissa illatiivi ilmaisee paikkaa johon mennään tai saavutaan. (MS, s)	a) Käyn suihkussa, pesen hampaani ja tupakoin. Sitten pukeun, suutelen tyttoani ja lähden ***yliopistossa***.
2) Illatiivimuodon käyttö inessiivimuodon asemasta paikkaa ilmaisevissa funktioissa. (MS, s)	b) Käyn ***suihkuun*** ja katson televisiota.
3) Sisäpaikallissijojen käyttö ulkopaikansijojen asemasta paikkaa ilmaisevissa funktioissa (MS, e)	c) Minun vuokra-asunnossa on yksi pieni huone. ***Tässä*** on myös kylpyhuone, vessa ja keittiö, mutta net on yhteiset.
4) Ulkopaikallissijojen käyttö sisäpaikallissijojen asemasta paikkaa ilmaisevissa funktioissa (MS, e)	d) Asun vuokra-asunnossa, kotini on ***toisella keroksella***.
5) Illatiivimuodon käyttö inessiivimuodon asemasta työtä tai toimintaa (3. inf.) ilmaisevissa funktioissa (MS, s)	e) …takia minulla oli syntymäpäiva ***tulemaan*** menin valitsen torttuja.
6) Sijapäätteen *-s* käyttö inessiivipäätteen *-ssA*	f) ***Yliopistos*** olen kaksitoistaksa neljaan.

Sisäpaikallissijojen norminvastaiset käyttötapaukset	
Käyttötapaus	**Esimerkki**
asemasta (M)	
7) Sijapäätteen *-sse* käyttö illatiivipäätteiden vokaali + *-n* , *-h-* + vokaali + *-n*, *-seen*, *- in*, *-hin* ja *-siin* asemasta (M)	g) Kun luennot ovat ohi, menen ***kahvilasse*** syömään…
8) Sijapäätteen *-ma* käyttö 3. infinitiivin illatiivipäätteen *-mAAn* asemasta (M)	h) ***Nukkuma*** menen oikein myöhään, välillä puoli kaksi.
9) Illatiivimuodossa puuttuu vokaali sanoissa, joiden illatiivi muodostetaan sijapäätteellä vokaali + *-n* (M)	i) Luennon jälkeen menen ***kirjaston*** tai ***postitoimiston***, jos tarvitse.
10) Sijapäätteen vokaali + *-n* käyttö illatiivipäätteen *-seen* asemasta (M)	j) Eteisesta saa komeroon, keittiöön, ***olohuoneen*** ja ***suihkuhuoneen***.
11) Verbirektion sijana (MS, e)	k) …harrastan aerobicia, uintia sekä ***lenkkeilystä.***
12) Elatiivimuoto inessivimuodon asemasta paikkaa ilmaisevissa funktioissa (MS, s)	l) Iltapäivisin käyn ***kirjastosta*** ja läksyt teken iltaisin voi aamuisin.
13) Inessiivimuoto elatiivin asemasta paikkaa ilmaisevissa funktioissa (MS, s)	m) Asun tällä hetkellä Tartossa asuntolassa, mutta olen kotoisin ***Tallinnassa***…
14) Taivutusvartalon virheet elatiivisijaisissa sanoissa (M)	n) …tykkään ***lämmistä*** ruuasta.
15) Taivutusvartalon virheet illatiivisijaisissa sanoissa (M)	o) Työlla olen kuudesta ***kahdeen***.
16) Genetiivimuoto inessiivimuodon asemasta paikkaa ilmaisevissa funktioissa (M, s)	p) ***Toisen nurkan*** on televisiota.
17) Illatiivimuoto	q) Tänän menen ostamaan

Sisäpaikallissijojen norminvastaiset käyttötapaukset	
Käyttötapaus	**Esimerkki**
nominatiivimuodon asemasta (F, s)	yhden ***valkosipuliin.***
18) Illatiivimuoto genetiivimuodon asemasta (F, s)	r) Koulun ***loppetamiseen*** jälkeen tuli minusta ***Tallinnaan yliopistoon*** ylioppilanen.
19) Vokaaliharmonian puuttuminen (F, e)	s) Vuoden ***vieressa*** kappin päällä on kynttilät ja sanakirja.

Corpus-based research of interior local cases used by estonian-speaking finnish learners

Keaty Siivelt

Summary

This article analyzes the use of interior local cases by estonian-speaking finnish learners on the basis of International Corpus of Learner Finnish (ICLFI). The aim of the paper is also to outline how L1 Estonian possibly influences the use of L2 Finnish.

The first part gives theoretical background to the reasearch and deals with theoretical issues of contrastive analysis, defining transfer, second language acquisition and the opportunities and methods of corpus based reasearch. In the second part, the data and methods are introduced. The data was collected from ICLFI (essays, compositions, term papers, short stories ect.) and contains of 36.636 words. In the third part the use of interior local cases are described and classified by functions. There are also presented the errors and the possible reasons of errors as well as proper forms. The paper concludes that morphological and phonological differences between L1 and L2 overcome the possible positive transfer on morfosyntactic level.

Esittely

BA Keaty Siivelt opiskelee maisteriopinnoissa Tallinnan yliopiston viron kielen ja kultuurin instituutissa, ksiivelt@tlu.ee

ALTERNATIIVSEID MOODUSEID FRASEOLOOGIA ESITAMISEKS SÕNASTIKUS

Katre Õim

Ülevaade

Artikli eesmärk on kirjeldada tuleviku fraseoloogiasõnaraamatu ülesehitust, mis ei eelda kasutajalt kuigi täpseid teadmisi fraseologismide leksikaalse koosseisu ja grammatilise struktuuri kohta. Võimalike valikmeetodite tutvustamisel toetun kognitiivse keeleteaduse uurimistulemustele. Analüüsitav keelematerjal pärineb Eesti kõnekäändude ja fraseologismide andmebaasist (EKFA)[1].

Võtmesõnad: kognitiivne keeleteadus, leksikaalne semantika, tesaurus, fraseoloogia[2]

[1] Baran, Anneli & Hussar, Anne & Õim, Asta & Õim, Katre (koost) 1998–2005. Eesti kõnekäändude ja fraseologismide andmebaas. http://www.folklore.ee/justkui (13.05.2009).

[2] Artikkel on seotud riikliku programmi „Eesti keele keeletehnoloogiline tugi (2006–2010)“ projektiga „Eesti fraseologismide elektroonilise alussõnastiku loomine“.

Sissejuhatus

Artikkel on ajendatud fraseologismide esitamisest eesti keele üld- ja erisõnastikes, mis arvestavad enamasti fraseologismide vormi, harvem sisu. Eesti väljendid antakse tavaliselt neis sisalduvate sõnade artiklite alamärksõnadena või näidete hulgas vastavalt sellele, kas sõnaühendi tähendust peetakse rohkem või vähem ülekantuks. Kuna enamik sõnastikke lähtub emakeelse kasutaja vajadustest, võidakse neis anda fraseologismi stiilimärgend, seletus ja variandid, kuid spetsiifilised grammatilised omadused tuleb osata näitelausetest välja lugeda. Ükskeelse sõnastiku vaikimisi eeldus on, et kasutaja valdab keelt, s.t tunneb mh väljendit, otsib tavaliselt sõnastikust lisa oma olemasolevatele teadmistele (Viks 2008: 252) – ja leiab püsiühendi vaevata ka artikli seest üles. „Eesti murrete sõnaraamatus“ suunavad suhteliselt idiomaatilise fraseologismi juurde selle kõik või olulised komponendid. „Fraseoloogiasõnaraamatus“ viivad fraseologismini selle komponendid, „Eesti kirjakeele seletussõnaraamatus“ ja „Eesti-vene sõnaraamatus“ fraseologismi suhteliselt vabalt valitud komponent, „Eesti keele mõistelises sõnaraamatus“, „Sünonüümisõnastikus“ ja „Väljendiraamatus“ fraseologismi tähendus ja leksikaalsed suhted.

Otsustamaks, millises sõnaartiklis võib fraseologism olla fikseeritud ja kirjeldatud, tuleb seega teada leksikaliseerunud väljendi täpset vormi, selles sisalduvaid sõnu (vt ka Espiñal 2005: 511) või tähendust. Nimetatud formaalseid tunnuseid pidi fraseologisme sõnaraamatuist pahatihti leida ei õnnestu. Muuhulgas võib seda takistada eesti fraseologismide suur süntaktiline paindlikkus ja leksikaalne muutlikkus (vt ka nt K. Õim 2005: 132). Ka võib fraseologism varieeruva esimese vm komponendi tõttu olla sõnaraamatus tähestikuliselt eri kohas. (K. Õim 2008)

1. Fraseoloogia metafooripõhine korraldamine

Kognitiivlingvistikas kirjeldatakse fraseologisme nende mõistestamisele toetudes tähenduse süstemaatilise mõistelise motivatsiooni ja konkreetsete metafoorsete või metonüümsete mõtlemisviiside alusel, mille väljendid aktiveerida võivad (Espiñal 2005: 520; vt ka Gibbs jt 1997). Kasutusel on termin *kehaliselt ja/või mõisteliselt motiveeritud tähendus*, s.t metafooride allik- ja sihtvaldkonna seos rajaneb maailma vahetul motoorsel kogemisel (näiteks korrelatsioonid motoorses kogemuses). Seejuures kuuluvad allikvaldkonda mõisted, mis tulenevad neist vahetutest kogemustest (vt Lakoff, Johnson 1980; vt ka Kövecses 2000a: 1, 2).

Tunnustatud kognitivist Zoltán Kövecses (2000a) on arutlenud fraseologismide korraldamise üle võõrkeele õppimise ja õpetamise kontekstis. Kövecsesi järgi aitaks fraseologismide omandamisele võõrkeeles palju kaasa nende mõisteline esitus. Meie mõistesüsteemis leiduvaga kooskõlas oleva süsteemi tõhusust kinnitavad õnnestunud testid. Niisiis tuleks õppijate jaoks ja väljendite mõistelisele motivatsioonile vastavalt osutada metafoori allik- ja sihtvaldkonnale (vt näidet 1) ning metonüümial rajanevate väljendite puhul valdkonnale, mida struktureerib mitmesuguste elementidega idealiseeritud kognitiivne mudel. (Kövecses 2000a: 3)

(1) allikala: ressurss (vt ka Lakoff & Johnson 1999: 161–164)

sihtala: inimene

INIMENE ON RESSURSS

*Na vanakese **jääse väega otsa**; Äi poiss tule tagasi, **sai otsa**; **Ega** ma **valla rahaga pole kasvatud**; Mul on kolm tütart ja kolm poega, kokku **pool vakamaad** lapsi.*

Selline esitusviis näitab, et idiomaatilised väljendid kuuluvad allik- ja sihtalaga kokku, s.t nende aluseks on mõistemetafoor. Mõistagi rakendatakse üht ja sama allikvaldkonda mitme sihtvaldkonna puhul (vt näidet 2) – need sihtvaldkonnad moodustavad metafoori ulatuspiirkonna (ingl *the scope of metaphor*) (vt ka Kövecses 2000b), mida tuleks samuti näidata.

(2a) INIMESE KEHAOSAD ON RESSURSS

Olen seda juba mitu-setu seitse korda ütelnud, ***lõuad on pähe ära kulunud****; Lähen käin turus,* ***ega jalad sinna ei jää****;* ***Ega*** *sul* ***pole kroonu jalad****, omi jalgu ei maksa ilmaaegu vaevata; Silmi* ***terve vakamaa*** *täis*.

(2b) ELUTÄHTIS ON RESSURSS

Tuleb ***justkui raha eest*** *seda vihma;* ***Ega*** *see* ***pole valla rahaga saadud****, et seda loobitakse; Ära nii palju söö,* ***ega*** *siin* ***ole mõisniku vara****.*

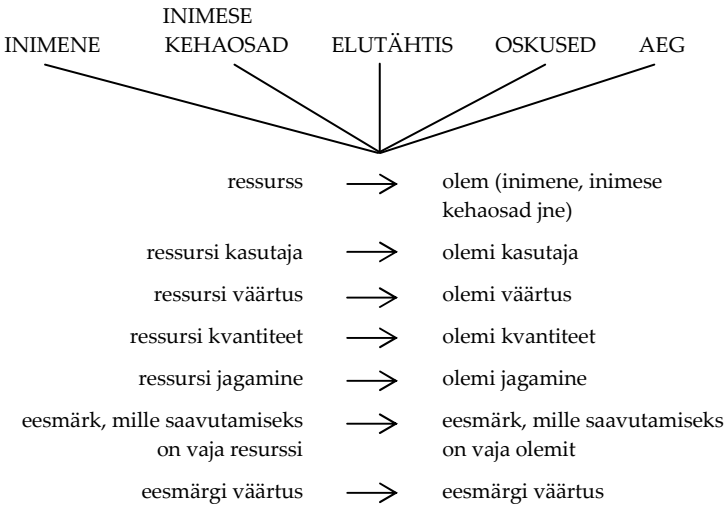
(2c) OSKUSED ON RESSURSS

Ei oska ***puud ega maad*** *joonistada.*

(2d) AEG ON RESSURSS

Aeg ***pole valla raha eest****.*

Niimoodi oleks meil oma mõistesüsteemi struktuuri tunduvalt lihtsam ette kujutada. Abstraktsete mõistete teatud aspektid justkui esindaksid vahetult teatud konkreetseid allikmõisteid, oleksid nende ülekantud variandid. Tavaliselt seatakse iga allikvaldkond ükshaaval vastamisi iga seda iseloomustava sihtalaga. Sedasi aga tuleb mitme sihtvaldkonna poolt jagatud allikvaldkond esile vaid korra (vt joonis 1) – koos kõigi sihtaladega, mille juures seda rakendatakse. (Kövecses 2000a: 7, 8)



Joonis 1. Allikvaldkond ressurss koos erinevate sihtvaldkondadega

Sellisest mõistesüsteemi metafoorse osa iseloomustamisest saab kasu olla vaid siis, kui täpsustatakse ka neid konkreetseid aspekte, mida allikvaldkond sihtvaldkondadele annab. (Kövecses 2000a: 8) Metafooril rajanevaid fraseologisme annab iseloomustada vähemalt kolme tähenduse alusel, mis sõltuvad allik- ja sihtala vahelistest olulistest ülekannetest.

Fraseologismide üldine tähendus tuleneb tõenäoliselt sellest, milliste sihtalade puhul konkreetset allikala rakendatakse ja kuidas väljendid on sõnastatud. Teades mis tahes metafooril rajanevat idiomaatilist väljendit ja sihtalasid, mille kohta selle allikala käib, teame ka fraseologismi potentsiaalseid üldisi tähendusi – sihtvaldkonnad ei saa jääda väljapoole allikvaldkonna ulatuspiirkonda. Näiteks ressursiga seotud väljendid saavad osa oma tähendusest sihtvaldkondadelt (isik, objekt, tegevus, protsess, aeg jne), millele see allikvaldkond rakendub.

Fraseologismide tunduvalt spetsiifilisem tähendus on ühelt poolt seotud metafoori lähteala struktuuriga, teiselt poolt sihtala vastava struktuuriga. Väljendite denotatiivne tähendus sõltubki igal üksikjuhul rakenduvast ontoloogilisest ülekandest, s.o allik- ja sihtala põhiliste koostiselementide vastavusest. Ressursi-metafoori moodustavaid ülekandeid kasutatakse kõikide sihtvaldkondade puhul, mis kuuluvad selle allikvaldkonna ulatuspiirkonda (vt joonis 1), s.t need ülekanded iseloomustavad võrdselt mõistemetafoore inimene on ressurss, aeg on ressurss jne. Seetõttu on fraseologismid, mida iseloomustab (kas ühe või mitme mõistemetafoori korral) sama ülekanne, ka sama tähendusega. (Kövecses 2000a: 9–12) Näiteks aitab ülekanne *ressursi väärtus → olemi väärtus* tuua selgust järgmiste väljendite tähendusse 'ilma vaevata [midagi] saama, saavutama' vms (vt näidet 3).

(3) *ega see mõisa vara ole, ega see pole valla rahaga saadud, ega see puu otsast lõigata ole, ei ole otsast võetav, nagu maast leitud, nagu taevast sadanud, näpuotsast võtta, varnast võtta*

Fraseologismide järelduslik ehk konnotatiivne tähendus sõltub episteemilisest ülekandest, mis toimub teadmuselt, mis kõnelejal on allikala elementide kohta, sihtala elementidele. S.t ühe mõisteala järeldusi, näiteks *kellegi kulul saadud ressurss on väärtusetu, seda võib raisata*, kasutatakse mõistemetafoori vahendusel teise mõisteala puhul. Taolised metafoorsed järeldused aitavad mh seletada sarnase denotatiivse tähendusega väljendite erinevaid tähendusi (vt näidet 4). (Kövecses 2000a: 12)

(4a) teiste oma või tehtut, niisama saadut võib raisata

Ega see *ometi* ***mõisniku vara ole****, mida sa nahka ajad;* ***Ega*** *ma* ***valla rahaga põle kasvatud****; Arvad, et see* ***puus on kasvanud****.*

(4b) vaevata saavutatu tuleb kasuks

See õnn on tal ***nagu maast leitud****; Niisugune koht* ***ei kuku ka taevast maha****.*

(4c) vaevata saavutatu ei ole midagi väärt

Ega *minagi* ***maast leitud ei ole****;* ***Ega*** *temagi* ***otsast võetud ei ole****, sel on visadust;* ***Ega*** *tüdruk* ***ei ole maast võtta ega puu otsast saada****.*

(4d) kasutamata ressurss on alati käepärast võtta

Ega *ma* ***varna otsas põle****, et sa võtad sealt; Kas ma* ***olen*** *talle siis* ***näpuotsast võtta****; Ega need jutud ole nii, et* ***võtad seina veerest****, andis tööd ka teha;* ***Ega*** *need head mehed sul* ***aia ääres vedele****; Mees või asi, tea, kust* ***aia äärest üles korjatud****; vrd* ***väljast korjatud*** *laps.*

Kui osutada metafoori allik- ja sihtalale ning ontoloogilistele ja epistemilistele ülekannetele nende valdkondade vahel, rikastab see fraseologismide mõistelist korraldamist ja seeläbi nende omandamist, aitab kätte näidata **fraseologismide tähenduse kujunemise ja muutumise.** Kirjeldatud esitusviisi rakendamist võib aga takistada töömahukus. Vaatamata sellele, et konkreetseid metafoori allikmõisteid leidub suure tõenäosusega tunduvalt vähem kui abstraktsemaid sihtmõisteid, tähendab metafooride ulatuspiirkonna määratlemine ka ainult fraseologismide põhjal põhjalikku etümoloogilist uurimistööd[3].

[3] Nii nagu Feliks Vakk (1970, 1984) seda Eestis ka teinud on.

2. Fraseoloogia metafoori sihtmõiste põhisest korraldamisest katalaani idioomide mõistelise sõnaraamatu näitel

Katalaani idioomide mõistelises sõnaraamatus „Diccionari de Sinònims de Frases Fetes“ (Espiñal 2004) on idioomid koondatud mõistelemmade ümber. Kokku sisaldab sõnaraamat 5500 tähestikuliselt reastatud mõisteartiklit, milles on 15 500 vastava mõistega semantiliselt või kognitiivselt seotud leksikaliseerunud väljendit. Materjalivalikul on lähtutud põhimõttest, et fraseologismid koosnevad mitmest sõnast ja on semantiliselt või (morfo)süntaktiliselt leksikaliseerunud. (Espiñal 2005: 509–513) Mõisteartiklite sees on väljendid järjestatud alfabeetiliselt ja esitatud sellises vormis, nagu nad esinevad enamikus sõnastiku koostamise aluseks olnud leksikograafilistest allikatest. Ligi kümnendik materjalist pärineb tänapäeva kirjalikest allikatest või suulisest kõnest. Vormiliste erisuste korral on antud väljend üldistatud kujul. (Espiñal 2005: 522, 539).

Selles kognitiivsest semantikast inspireeritud sõnaraamatus ei eeldata, et kasutaja juba teab iga konkreetset väljendit, pigem ollakse seisukohal, et kasutaja teadmised fraseologismist sõltuvad suuresti võimest tajuda mõistelisi seoseid. Seega pakutakse kasutajale hulgaliselt mõiste ja väljendi seoseid, mille abil jõuda millise tahes väljendini. Katalaani idioomide mõistelises sõnaraamatus ei peeta fraseologisme mitte spetsiifilisteks sõnaühenditeks, mis kuuluvad keele sõnavarasse ja on juhuslikult seotud teatud märksõnadega, vaid neid käsitatakse süntaktiliste tarinditena, mis sõltuvad mõistevaldkondadest kui erinevate kognitiivsete protsesside (s.o metafoor, metonüümia jm allikas – sihtülekanded) kaastulemid (vt ka Gibbs jt 1997: 142). Sõnaraamatu mõisteloend on kujunenud põhiliselt nende mõisteliste sihtvaldkondade põhjal, millega fraseolo-

gismid inimlikus mõistesüsteemis seostuvad; väiksem osakaal on metafoori allikala struktuurist tulenevatel mõistetel. (Espiñal 2005: 513, 523) Väga ebamäärastest ja ülekantud mõistetest on hoidutud; kahe või enama mõistega on seotud polüseemilised väljendid (vrd eesti väljendit *hunnik õnnetust* – saamatu, õnnetu), või sellised fraseologismid, mille tähendus seostub rohkem kui ühe mõistelise sihtalaga.

Sõnaraamatu mikrostruktuur on järgmine. Kõigepealt esitatakse väljendi grammatiline kategooria (mis on enamasti korrelatsioonis vastava mõiste grammatilise kategooriaga), seejärel antakse fraseologismi leksikograafiline definitsioon ja näide koos viitega leksikograafilisele allikale. Vajaduse korral kommenteeritakse väljendi kasutust. Samuti viidatakse fraseologismidevahelistele mõistelistele suhetele. Kui vaja, täpsustatakse formaalseid variante, antakse murdevariandid ja murdemarkeering, morfoloogilist (liitsõnade ja reduplikatiivväljendite kohta) ja normatiivset infot (selliste sõnade kohta, mida tänapäeva normatiivsetest sõnaraamatutest ei leia), selgitatakse väljendite mõnede koostiselementide etümoloogiat. Väljendi peasõna järgi määratletakse selle süntaktiline kategooria: vahet tehakse nimisõnafraasidel, mille tuumaks on nimisõna, hulgafraasidel, mille tuumaks on hulgasõna jne. (Espiñal 2005: 514 jj, 526)

Lisaks mõistelisele osale sisaldab nii elektrooniliselt kui ka paberkandjal ilmunud sõnaraamat kõigi käsitletavate väljendite alfabeetilist indeksit koos viidetega vastavatele mõistetele. (Espiñal 2005: 509, 513)

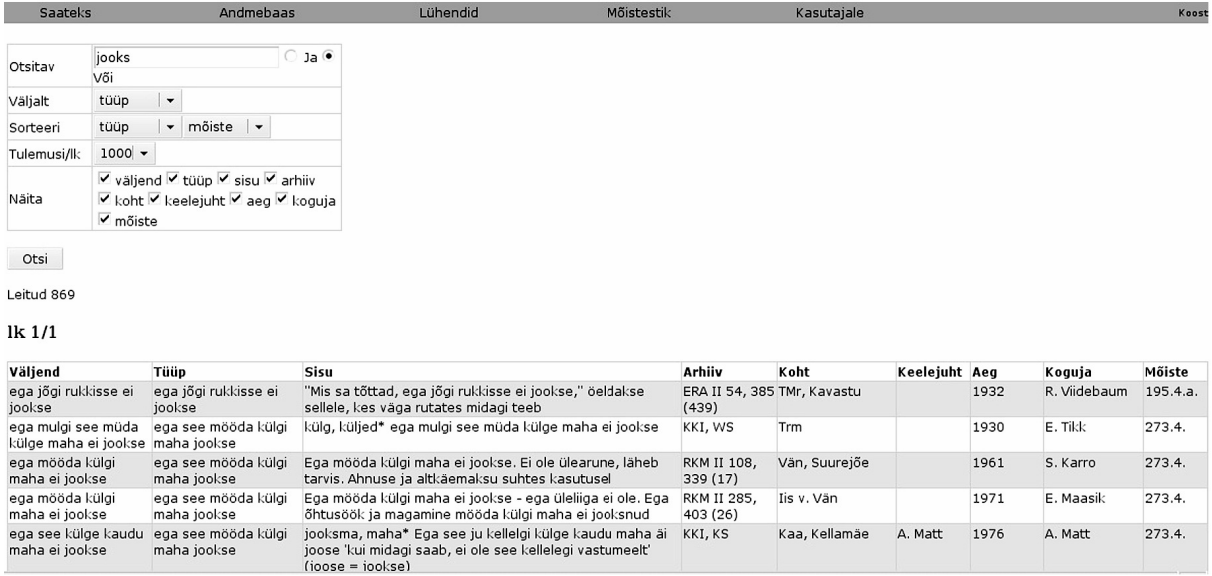
3. Eesti fraseoloogia metafoori sihtmõiste põhine korraldamine

Seni ilmunud eesti keele sõnaraamatutest on fraseologismid mõistepõhiselt korraldatud „Väljendiraamatus". Põhimõtteliselt on sama lähenemisviisi võimalik tulevikuski rakendada ja vormistada näiteks eesti kõnekäändude ja fraseologismide andmebaasi materjali põhjal mahukam ja detailsem mõisteline fraseoloogiasõnaraamat. See on ka riikliku programmi „Eesti keele keeletehnoloogiline tugi" projekti „Eesti fraseologismide elektroonilise alussõnastiku loomine" põhiline eesmärk. Täpsemalt plaanitakse toota eesti fraseologismide mõisteline alussõnastik elektroonsel kujul, milles on eri mõistetega seostuvad fraseologismid koondatud mõisteartiklitesse ning milles antakse teavet fraseologismide morfo-süntaktilise ehituse ja kasutuse kohta loomulikus keeles (selle võimalustest ja seostuvatest probleemidest vt lähemalt näiteks Moon 2008).

3.1. Eesti fraseologismide elektroonilise alussõnastiku põhi ja makrostruktuur

Alussõnastikul on kaks põhilist toetuspunkti: 1) Eesti kõnekäändude ja fraseologismide andmebaas, mis sisaldab 154 549 arhiiviteksti, s.o ligikaudu 35 000 väljendit (vt joonist 2); 2) Eesti kõnekäändude ja fraseologismide mõistestik[4], mis koondab endas umbes 5000 alam- ja 2500 põhitasandi mõistet (vt joonist 3).

[4] Õim, A. (koost), Õim, K. (toim) 2008. Eesti kõnekäändude ja fraseologismide mõistestik. http://www.folklore.ee/justkui/moiste.php (13.05.2009).



Joonis 2. Eesti kõnekäändude ja fraseologismide andmebaas

1.1	**aadel *sinine veri***	Vaata
2.3	**abielu**	Vaata
2.3.1	mitte abielus *süda on vaba*	Vaata
2.3.2	abieluvoodi *abielu sepapada*	Vaata
2.3.3	õnnetu abielu *ihu on rikutud ja hing on hukatud*	Vaata
2.4	**abielluma *paari minema***	Vaata
2.4.1	abieluettepanekut tegema *kätt paluma*	Vaata
2.4.2	abieluettepanekut vastu võtma *oma kätt ja südant andma*	Vaata
2.4.3	abielluda kavatsema *plaani pidama*	Vaata
2.4.4	abiellumast keelduma	Vaata
2.4.5	iga hinna eest abielluma	Vaata
2.4.6	enne vanemat õde või venda abielluma *ajab seanaha selga*	Vaata
2.4.7	naise seisukohast (meest võtma) *päitseid pähe ajama*	Vaata
2.4.8	naist võtma *altari ette viima*	Vaata
2.4.9	abielus olema *rõngas on ninas*	Vaata
2.4.11	vabaabielus olema *armukest elama*	Vaata
2.4.12	abiellumises kokku leppima	Vaata
2.4.13	abielluda enam mitte tahtma	Vaata
2.4.14	ilma lasteta abielu elama *neitsipõlve elama* NB! Vt lastetu nr 262.6	Vaata
2.4.15	endast vanema naisega abielluma *kellelgi on sea õnn*	Vaata
2.4.16	abiellumisest teatama (kirikus), kihluma *maha kuulutama*	Vaata
2.4.17	mehele panema, abielluma sundima	Vaata
2.4.18	mehele saama	Vaata
2.4.19	mitte abielluma, katki jääma (abiellumise kohta)	Vaata
2.4.20	väga noorelt abielluma	Vaata
2.4.21	lesega abielluma	Vaata
2.5	**abielus olnud (lahutatud, leseks jäänud)**	Vaata

Joonis 3. Eesti kõnekäändude ja fraseologismide mõistestik

Alussõnastiku tarbeks on EKFAst välja valitud 20 671 tüpoloogiliselt, morfoloogilis-süntaktiliselt ja semantiliselt märgendatud fraseologismi. Korpuste mahu erinevus tuleneb põhiliselt sellest, et EKFA sisaldab hulgaliselt ainukordse kujundkõne näiteid, väljend- ja ühendverbe jm mittefraseoloogilisi ja/või mittekinnistunud keelendeid, millest paljud on üles kirjutatud vaid paaril korral.

EKFA põhitasandi mõistete põhjal on sõnastiku jaoks formaliseeritud 998 ülemtasandi mõistet. Viimaste alusel tekkivad mõisteartiklid ongi Eesti fraseologismide elektroonilise alussõnastiku põhiüksused. Kuna sõnavarale omaselt on erinevad semantilised väljad kaetud keelematerjaliga väga ebaühtlaselt,

siis suurim fraseologismide arv ühes mõisteartiklis on ligikaudu 50 ja vähim 1. Osaliselt võib sellise tulemuse tingida EKFA materjali piiratus.

Seega osutatakse sõnastikus mõistetasandil metafoori sihtvaldkondadele, mis motiveerivad koos allikvaldkondadega suuremat osa fraseologismidest, ja viimasest tulenevalt ka fraseologismide üldisele tähendusele. Mõiste kajastab olendeid, esemeid ja nähtusi oluliste ja spetsiifiliste tunnuste, seoste ja suhete kaudu. Nagu öeldud, võib mõistete maht olla väga erinev, näiteks mõistes *peksma* eristuvad alamõisted vastavalt peksmise intensiivsusele ja peksmisvahendile jne, mõistes *sööma* vastavalt söömise kiirusele ja toidu kogusele, mõistes *laiskus* vastavalt omaduse määrale. Piisavalt ulatuslike mõistete puhul on tõenäosus, et kasutaja suudab neid väljenditega seostada, suurem. Mõistagi peaks sõnaraamatus fraseologismidega kaasas käima info nende vormi, tähenduse ja kasutuse kohta (vt ka Espiñal 2005: 513). Kuna eesti fraseologisme ei ole sõnaraamatutes (v.a „Fraseoloogiasõnaraamat" ja „Väljendiraamat") kuigi süsteemselt käsitletud, pole paraku mõeldav anda iga väljendi juures selle leksikograafilist definitsiooni.

Küll viib loodud mõistesüsteem kokku semantiliselt seotud väljendid, aitab demonstreerida fraseologismide tähenduse arenemisjärke ja teisenemist, funktsioonimuutusi (vt allpool näidet 5 ja tabelit 3). Nii ühes mõisteartiklis kui ka eraldiseisvates artiklites seob sama üldise tähendusega väljendeid põhinemine samal allikmõistel, näiteks *kont kõhus*, *kange kui pulk*, *kange kui nui*, *nagu küünarpuu alla neelanud*. Sama üldise ja denotatiivse tähendusega väljendeid, mis toetuvad samale ontoloogilisele ülekandele, võib pidada sünonüümideks (vt tabel 1), näiteks *kont kõhus*, *nagu küünarpuu alla neelanud.* Sama üldise, denotatiivse ja konnotatiivse tähendusega väljendid on täissünonüümsed, sest toetuvad samale episteemilisele ülekan-

dele, näiteks *kange nagu puuhobune*, *nagu puukaru*. Katalaani idioomide mõistelises sõnaraamatus on suurt tähelepanu pööratud idioomide sünonüümiasuhetele, lisaks on teatud juhtudel näidatud veel antonüümiasuhe, pöördvastandus või argumendi pöördjärg ja laiendatud tähendus (põhjus–tagajärg, protsess–seisund, tegevus–moodus, erinevad tähendusvarjundid) (Espiñal 2005: 528). EKFAs on ristviitamist seni rakendatud vaid leksikaalselt äärmiselt lähedaste või sünonüümsete väljendite vahel ja seda mitte kuigi järjekindlalt. Küll tagab andmebaasi märgendustase selle, et samasse mõisteartiklisse koonduvad nii sünonüümid kui ka antonüümid (näiteks kasvama: *ajab kasu taga – ei kosu ega kasva*; peavari: *varju all olema – aia ja hange vahel olema*; kiire: *pole aega pihku sülitada – aega küll selle kiire asjaga*) ja osa-tervik suhetes fraseologismid (näiteks leib: *kurakäemees*; *peremees*; *näljaleib*; *jõuluorikas*; *kahe käe leib*; *saksaviilukas*; *sandikäär*; haigus: *Tooma kange käsi*; *veiseröögatus*; *ei saa istu ega astu*; *tõbi suust sisse läinud*; *küüneviha sisse läinud*; laps: *harkjalg*; *kisapill*; *peenike pere*; *silmatera*; *jalapäästja*; *kaapekakk*; *peasamuna*; *kasukalaps*; *kõrvaline laps.*

Tabel 1. Fraseologismide leksikaalseid seoseid

Tähendus	**Mõisteline seos**	**Sünonüümia**	**Täielik sünonüümia**	**Antonüümia**	**Osatervik**
sama üldine tähendus	+	+	+	+	+
sama denotatiivne tähendus	–	+	+	–	–
sama konnotatiivne tähendus	–	–	+	–	–

Samuti on EKFAs rohkesti jälgi väljendite tähenduse üldistumisest ja seda kõigil kolmel tähendustasandil. Vrd väljendite *puud ja maad*, *maid jagama*, *puile ega maile* kasutusi näites 5.

(5a) *See on hakkaja mees, see saab omale varsti* ***puud ja maad*** 'saab omale varsti majapidamise'; *Kõik on ta ära joonud,* ***puud ja maad*** *ära priisanud, terve varanduse;* ***Ei saa puile ega maile*** 'ei saa omas ametis edasi'; *Alati tagasipite,* ***ei saa puele ei maele*** 'inimene ei sua oma harilikku leiba, ei riiet'.

(5b) ***Mis puid-maid*** *teil ka* ***jagada on****, et te korda ei saa?* ***Mis ossa vai puu-maa*** *teil* ***ütte putusse*** *kui ütstõsega iluste läbi ei saa? Jüri ja Jaan said poe juures kokku ja läksid kohe* ***maid jagama****;* ***Mis talud*** *teil* ***jagada on****?* ***Mis metsad-maad*** *neil* ***kokku puutuvad****, et tülli lähevad?* ***Mis mõisa*** *neil* ***jaka om vai mõisakraam****?*

(5c) *Ta sööb* ***puu ja maa*** *kokku; Ei oska* ***puud ega maad*** *joonistada; Kus see kasvab, kas* ***puis või mais****? Oled sa kasvanud* ***puis või mais****?*

(5d) *Nüüd ei saa temaga enam* ***puile ega maile*** 'ei saa sugugi läbi'; *Räägi* ***puile ehk maile*** 'kui laps ei kuula sõna'.

3.2. Eesti fraseologismide elektroonilise alussõnastiku mikrostruktuur. Mõisteartikli osad

Fraseologismide järjestamiseks mõisteartikli sees teeme kõigepealt vahet 1) fraseologismidel, mis väljendavad lausega tähistatava situatsiooni komponente, ja 2) lausekujulistel fraseologismidel. Edasi võtame arvesse seda, et lause moodustajate hulgas on eri tüüpi fraase, ja lausetel-fraseologismidel on eri suhtluseesmärgid (vt tabelit 2).

Tabel 2. Fraseologismide funktsioon ja ehitus

Tüüp	**Jaotus**	**Näide**
	TEGUSÕNAFRAASID	*ei karda vanakuraditki*
	NIMISÕNAFRAASID	*ausõna aukudega; paljas luu ja nahk*
LAUSE MOODUS-TAJAD	OMADUSSÕNAFRAASID	*lollim kui siga*
	MÄÄRSÕNAFRAASID	*mehemoodi; läbi ja lõhki*
	KAASSÕNAFRAASID	*oma käe peal*
	HULGASÕNAFRAASID	*paar parajaid*
	VÄITLAUSED	*Kopik läheb enne peos palavaks, kui välja annab.*
	KÜSILAUSED	*Kas tahad suuremaks kasvada?*
LAUSED	KÄSKLAUSED	*Hoia oma leivamulk kinni!*
	SOOVLAUSED	*Hääd üüd ja kirbul tüüd!*
	HÜÜDLAUSED	*Vaat, kus mul asjamees!*
	MITMEST LAUSEST SÕNAÜHENDID	*Kust tuled? – Sinna, kuhu läksin.*

Fraseologismide väga esialgne märgendamine funktsiooni ja ehituse järgi näitab, et tegusõnafraase on rohkem kui 20 000 fraseologismi hulgas umbes ⅓, nagu ka nimisõnafraase ja lausekujulisi väljendeid. Omadus-, määr-, kaas- ja hulgasõnafraasid moodustavad materjalist kokku vähem kui kümnendiku. Verbita lauselühendite ja elliptiliste lausete eristamisel

tekib praegu veel rohkesti küsitavusi. Edasise analüüsi käigus need vahekorrad tõenäoliselt mõnevõrra muutuvad.

Kui suurem osa EKFA materjalist on märgendatud mh vastavalt sellele, millise sõnaliigiga on fraseologism korrelatsioonis (tegusõnalised väljendid vastavad küsimusele *mida teeb?*, omadussõnalised väljendid küsimusele *missugune?*, kvantumit tähistavad nn hulgasõnalised väljendid küsimusele *kui palju?* jne), siis alussõnastikus on sellest põhimõttest loobutud. Seda põhiliselt nii EKFA materjali sisulise kui ka vormilise ebamääruse ja laialivalguvuse tõttu (vt näidet 6) .

(6) vrd *Kiidus mees **nagu kukk** teeb kanakarja ees kok-kok; Keksib **nagu kukk**; Istus **kui kukk**, liiguta-i lillegi; Oled **justkui** hollandi **kukk**; Kõnnid uhkelt paneeli peal **kui kukk** aia otsas; Ära ole nõnda kahmakas ja upsakas oma sõnudega, kange ütlemäie, egä sa viel **kui kikkas** et õle; Siis sõidab hobu **kut kukk**, mies peal kut nukk*.

Arusaadavalt ei pruugi fraseologismi kui terviku sõnaliigimääratlus olla vastavuses tema fraasitüübiga. Näiteks nimisõnafraas *tuli takus* võib lauses esineda nii omadus- kui ka määrsõnalisena, vrd *Mes rumala sa alade tied, et aeva on tuli taga?* ja *Jookseb nii et kas mu pärast, nagu oleks tuli takus*.

Fraseoloogiliste käsk-, soov- ja hüüdlausete hulgas on hellitusnimesid, sõimusõnu ja pilkeid (*kukumuna*, *täinahk*, *mesimeeleke*, *kuriloom*), hüüatusi ja hüüdeid (*Tule taevas appi!*), viisakusvormeleid ja soove (*Õnn kaasa!*), sajatusi (*Susi sind söögu!*) jm. Tihti kiilduvad need fraseologismid ümbritsevasse lausesse, jäädes viimasega grammatiliselt seostamata – sagedased on nominaal- ja hüüdfraaslaused, s.o mitteverbaalsest fraasist koosnevad vaeglaused (vt EKG II: 102, 228).

Mitmest (osa)lausest koosnevad sõnaühendid jäävad alussõnastiku materjali hulgas oma tunnuste poolest äärealale. Sedalaadi ehitusega ütluste hulgas on tõrjevormeleid, paree-

rivaid repliike ja sõnamõnitusi (näiteks *Imelik. – Imelik oli Tootsi koolivend*), kiirkõnesid (*Õtsekõhe Õtsa Mardi õuest läbi*), sõnamänge (*„Kuradi vunts," ütles habe*), lõppriimilisi ütlusi ja salme (*Eit läks heina, lõi sarved seina*) jne. Paljud taolised väljendid paistavad keskenduvat verbaalsusele. On loomulik, et lausekujulised fraseologismid ei ole morfosüntaktiliselt nii kinnistunud kui fraasid. Mida paremini väljend semantiliselt jaguneb, seda tõenäolisem on selle süntaktiline paindlikkus (vt Espinal 2005: 520; väite poolt- ja vastuargumentidest vt eesti keeles Muischnek 2006: 15, 16).

Fraseologismide jagamisele funktsiooni ja ehituse põhjal järgneb nende morfoloogiline liigitus. Lisa 1 (vt allpool) sisaldabki metainfot näiteväljenditesse kuuluvate sõnade morfoloogilise analüüsi tulemuste kohta. Tõenäoliselt tuleb fraseologismide morfoloogilise analüüsi tulemused ühestada käsitsi – väljendites pole automaatseks morfoloogiliseks ühestamiseks lihtsalt piisavalt lausekonteksti. Lisas 1 on välja toodud ka fraseologismide kõige iseloomulikum morfosüntaktiline tunnus (fraasi peasõna kõrval) – ühtlasi moodustab see sama fraasi- või lausetüüpi jagavate väljendite morfo-süntaktilise ühisosa.

Fraseologismide levikuandmed esitatakse alussõnastikus kihelkonna täpsusega; fraseologismide esinemissagedust ei näidata, näiteid levikuga ei seota. Selle põhjuseks on eelkõige EKFA andmete ebaühtlus, piiratus ja fraseoloogiasõnastiku spetsiifika.

Nagu eespool toodud rohketest näidetest selgub, näitab fraseologismide tuumelemendi varieerumine grammatilisi ja leksikaalseid mooduseid, kuidas on kasutatud üht või teist kujundit. Kõiki neid teisendeid aga ei ole sugugi õige pidada leksikaliseerunuks (probleemist lähemalt vt nt K. Õim 2005, K. Õim jt 2003: 11 jj). Et mitte piirata kõikvõimalike väljendivariantide esitamisega fraseologismide kasutamist loomulikus

keeles, tutvustame nende vaheldumist alussõnastiku mõisteartikli näiteplokis. Seejuures valime EKFA üleskirjutuste hulgast sellised, mis toovad esile ja täpsustavad võimalikult palju ja erinevat grammatilist infot fraseologismide kohta – lähtume fraseologismide morfosüntaktilisest, süntaktilisest ja leksikaalsest varieerumisest fraseologismi-lemma suhtes. Fraseologismide suhteliselt vaba kasutuse tõttu eesti keeles tuleb ühe väljendi kohta anda tõenäoliselt suhteliselt suur hulk erinevat laadi näiteid. Sobiva materjali puudumisel näiteid aga ei esitata. Näiteplokis toome ära ka tüüpilised murdevariandid; vajaduse korral seletame murdekeelendite tähendust. Fraseologismide semantiline varieerumine, sh polüseemia, tähendusnihked, kajastub juba loodud mõistesüsteemis.

Seega kujundatakse metafoori sihtmõiste põhised sõnastikuartiklid mikrotasandil mitme väga erineva parameetri alusel. Lõpptulemusena esitatakse eesti fraseologismide elektroonilises alussõnastikus üheskoos sama või sarnase tähendusega fraseologismid, millel on sama või sarnane ehitus.

Kokkuvõte

Eesti fraseoloogia esitatakse sõnastike sihtgrupist ja otstarbest olenevalt sõnaraamatute makro- või mikrostruktuuris. Et otsustada, millises sõnaartiklis võib fraseologism olla fikseeritud ja kirjeldatud, tuleb teada leksikaliseerunud väljendi täpset vormi, selles sisalduvaid sõnu või tähendust. Arvestades eesti fraseologismide suurt süntaktilist paindlikkust ja leksikaalset muutlikkust, ei ole see alati paraku võimalik.

Fraseologismide esitamisele formaalsete tunnuste põhjal pakuvad alternatiivi kognitiivsest keeleteadusest inspireeritud metafooripõhised lähenemised, s.t fraseologismid esitatakse erisõnastikus mõisteseoste alusel. Väljendite mõistelisest motivat-

sioonist tulenevalt sõltub meie teadmus väljendi kohta suure tõenäosusega võimest tajuda mõistelisi seoseid. Mõiste kajastab olendeid, esemeid ja nähtusi oluliste ja spetsiifiliste tunnuste, seoste ja suhete kaudu. On ilmne, et piisavalt mahukaid mõisteid suudaks sõnaraamatu kasutaja väljenditega seostada. Fraseologismide mõistelist korraldamist rikastaks osutamine metafoori allik- ja sihtalale ning ontoloogilistele ja episteemilistele ülekannetele nende valdkondade vahel. Selle kaudu oleks fraseoloogiasõnastikus mh võimalik kätte näidata fraseologismide tähenduse kujunemine ja muutumine. Vaatamata sellele, et konkreetseid metafoori allikmõisteid leidub suure tõenäosusega tunduvalt vähem kui abstraktsemaid sihtmõisteid, tähendab metafooride ulatuspiirkonna määratlemine siiski põhjalikku etümoloogilist uurimistööd. Katalaani idioomide mõistelises sõnaraamatus ei peetagi fraseologisme spetsiifilisteks sõnaühenditeks, mis kuuluvad keele sõnavarasse ja on juhuslikult seotud teatud märksõnadega, vaid neid käsitatakse süntaktiliste tarinditena, mis sõltuvad mõistevaldkondadest kui erinevate kognitiivsete protsesside kaastulemid.

Ka eesti kõnekäändude ja fraseologismide andmebaasi materjali põhjal koostatavas eesti fraseologismide elektroonilises alussõnastikus koondatakse eri mõistetega seostuvad fraseologismid mõisteartiklitesse ning antakse teavet fraseologismide morfo-süntaktilise ehituse ja kasutuse kohta loomulikus keeles. Mõisteartikli sees tehakse vahet fraseologismidel, mis väljendavad lausega tähistatava situatsiooni komponente, ja lausekujulistel fraseologismidel. Edasi arvestatakse seda, et lause moodustajate hulgas on eri tüüpi fraase ja lausetel-fraseologismidel on eri suhtluseesmärgid. Fraseologismide jagamisele funktsiooni ja ehituse põhjal järgneb nende morfoloogiline liigitus. Fraseologismide vaheldumise demonstreerimisel alussõnastiku mõisteartikli näiteplokis lähtutakse fraseologismide morfosüntaktilisest, süntaktilisest ja leksikaalsest varieerumi-

sest fraseologismi-lemma suhtes. Niisiis kujundatakse metafoori sihtmõiste põhised sõnastikuartiklid mikrotasandil mitme väga erineva parameetri alusel. Lõpptulemusena esitatakse eesti fraseologismide elektroonilises alussõnastikus üheskoos sama või sarnase tähendusega fraseologismid, millel on sama või sarnane ehitus.

Edaspidi oleks alussõnastiku tarbeks morfoloogiliselt töödeldud fraseologisme võimalik analüüsida süntaktiliselt fraasistruktuuri või moodustajate süntaktiliste funktsioonide järgi, pidades silmas perspektiivi alustada fraseologismide struktuurimustrite ja ülekantud tähenduse vastavuste tuvastamist.

Kirjandus

Eesti kirjakeele seletussõnaraamat I–VII. Tallinn: Eesti Keele Instituut, 1988–2007.

Eesti murrete sõnaraamat I–lahhest. Tallinn: Eesti Keele Instituut, 1994–2007.

Eesti-vene sõnaraamat 1–4. Tallinn: Eesti Keele Sihtasutus, 1997–2006.

Espiñal, Maria Teresa 2004. Diccionari de sinònims de frases fetes. Barcelona / València: Universitat Autonòma de Barcelona. Servei de Publicacions / Publicacions de la Universitat de València / Publicatcions de l'Abadia de Montserrat.

Espiñal, Maria Teresa 2005. A conceptual dictionary of catalan idioms. – International Journal of Lexicography. Vol. 18, No. 4, 509–540.

Gibbs, R. jt 1997 = Raymond W. Gibbs, Jr. & Josephine M. Bogdanovich, Jeffrey R. Sykes & Dale J. Barr 1997. Metaphor in Idiom Comprehension. – Journal of Memory and Language, 37, 141–154.

Kövecses, Zoltán 2000a. A cognitive linguistic view of learning idioms in an FLT context. Essen: LAUD Linguistic Agency, University-GH Essen.

Kövecses, Zoltán 2000b. The scope of metaphor. – Metaphor and Metonymy at the Crossroads: A Cognitive Perspective / Ed. by A. Barcelona. Berlin–New York: Mouton de Gruyter, 79–92.

Lakoff, George & Johnson, Mark 1980. Metaphors We Live By. Chicago and London: The University of Chicago Press.

Lakoff, George & Johnson, Mark 1999. Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought. New York: Basic Books.

Moon, Rosamund 2008. Dictionaries and collocations. – Phraseology: An interdisciplinary perspective / Ed. by S. Granger, F. Meunier. Amsterdam/Philadelphia: John Benjamins Pub. Co, 314–336.

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

Saareste, Andrus 1958–1963. Eesti keele mõisteline sõnaraamat I–IV. Stockholm: Vaba Eesti.

Vakk, Feliks 1970. Suured ninad murdsid päid ... (Pea ja selle osad rahvalike ütluste peeglis). Tallinn: Valgus.

Vakk, Feliks 1984. Miks just nõnda? Peotäis tekkelugusid ja uudistavaid lühimatku eesti fraseoloogia radadelt. Tallinn: Valgus.

Viks, Ülle 2008. Eesti-X-keele sõnastik ja grammatika. – Eesti Rakenduslingvistika Ühingu aastaraamat 4. = Estonian Papers in Applied Linguistics 4 / Toim. H. Metslang, M. Langemets, M.-M. Sepper. Tallinn: Eesti Keele Sihtasutus, 247–261.

Õim, Asta 1993. Fraseoloogiasõnaraamat. Tallinn: Eesti Keele Sihtasutus.

Õim, Asta 1998. Väljendiraamat. Tallinn: Eesti Keele Instituut.

Õim, Asta 2007. Sünonüümisõnastik. Tallinn.

Õim, K. jt 2003 = Katre Õim & Asta Õim & Anne Hussar & Anneli Baran 2003. Kõnekäändude kartoteek andmebaasiks. – Keel ja Kirjandus, 1, 4–23.

Õim, Katre 2005. Fraseologism *versus* kõnekäänd. – Emakeele Seltsi aastaraamat 50 / Toim. M. Erelt. Tallinn: Emakeele Selts, 129–142.

Õim, Katre 2008. Fraseoloogia ja sõnaraamatud. – Keel ja Kirjandus, 10, 774–789.

Lisa 1. Näide mõisteartiklist AITAMA, ABI

Fraasi-/ Lausetüüp	**Fraseologism**	**Morfoloogiliselt analüüsitud fraseologism**	**Morfo-süntaktiline ühisosa järgmise väljendiga**	**Jälg**	**Näide**
NIMI-SÕNA-FRAAS	**armuadra-mees**	armuadramees armu_adra_mees+0 //_S_ sg n //	Liitsõna	Emm, Han, Khk	Mis armuadramees ta mool eige on, et ma pea teda ühtelugu aitama; Mis armuadrapoiss sa oled, et meitega ühüpailu palka tahad. Ei saa, ikka töö järel saame; Mis armuader see mol on; Mis armukott sa mool oled.
	hädа-ankur	hädaankur häda_ankur+0 //_S_ sg n //		Kuu	Oled toistele hädäankuriks old.

	tugi ja abi	tugi tugi+0 //_S_ sg n // ja ja+0 //_J_ // abi abi+0 //_S_ sg n //	rinnastus	Juu, Kod, Koe, Kse	Eks pueg ole mo tugi ja abi; Isä abi, emä tugi, nõna üeldase veikes last; Eks ta ole oma emal abiks ja tueks.
	ei abi ega armu	ei ei+0 //_V_ neg // abi abi+0 //_S_ sg p // ega ega+0 //_J_ // armu arm+0 //_S_ sg p //		Hls, Hää, Jõh, Kaa, Krk, Kuu, Lüg, Saa, Vil	Es saa selelst suigunuiast mingisugust abi ega armu; Asi oo küll sõhuke, et kas vei karju abi, aga mis seegid aitab, abi-armu pole kohegilt tulemas üht; Mea ole ikki sii ostet ori, miul ei ole hoolekandjet ega armuandjet.
	parem käsi	parem parem+0 //_C_ sg n // käsi käsi+0 //_S_ sg n //	eestäiend	Kaa, Mar, Tür	Mo param käsi läks ää kõrvast; Abar oli meistri parem käsi ja vasak jalg ning mees igale poole varnast võtta.

	nagu teine käsi	nagu nagu+0 //_J_ // teine teine+0 //_O_ sg n // käsi käsi+0 //_S_ sg n //	võrdlus-tarind		Sa oled moole na kut teina käsi; Köögin om turhslaki ninda vaja nagu tõist kätt; Puudub nagu teine käsi; Sjõ lats om mul kui uma käsi.
	kui käsi-kannel	kui kui+0 //_J_ // käsikannel käsi_kannel+0 //_S_ sg n //		Hää, Saa	Teisele nagu käsikannel olema.
TEGU-SÕNA-FRAAS	**saab käed pikemaks**	saab saa+b //_V_ b // käed käsi+d //_S_ pl n // pikemaks pikem+ks //_C_ sg tr //	sihitis	Rei, Se, Vas	Tüdruk oleks juba kätt pikemaks olnd; Mul ka latsõkõnõ jo pikep kätt; Täl om ju käsi pikk. Vanembast, kellel latse juba hää käskjala omma.
	auku täitma	auku auk+0 //_S_ sg p // täitma täit+ma //_V_ ma //	sihitis	Mar	Tä täidäb ikka mõne augu ää jo.

	jalatäisi lühendama	jalatäisi jala_täis+i //_S_ pl p // lühendama lühenda+ma //_V_ ma //		Har, Juu, Koe, Kuu, Mar, Muh, Räp	Minge aage luumad ää, lühönda mu jalga; Löhendasid mu sammusi koa, mul põle nii pailu käimist alati; Jätä mu jalasammu tagasi, vii mu vellele tiid, et ta hommõn mu poolõ tulõ; Küll ma ta jalad lühendän.
	otsib abi aiast, toetust tugiteibast	otsib otsi+b //_V_ b // abi abi+0 //_S_ sg p // aiast aed+st //_S_ sg el // toetust toetus+t //_S_ sg p // tugiteibast tugi_teivas+st //_S_ sg el //	määrus	Kad, Vas, Vil, Se	Ajateivas on su abi ja toeteivas on su tugi; Tema abi aias, tuetus toeteibas.
	joone peale aitama	joone joon+0 //_S_ sg g // peale peale+0 //_K_ // aitama aita+ma //_V_ ma //	määrus	Kjn, Koe	Joone peale aitama, aitama paremusele; Joone piäle säädmä.
	teiste õlul elama	teiste teine+te //_P_ pl g // õlul õlg+ul //_S_ pl ad // elama ela+ma //_V_ ma //		Juu, Kär, Mär	Söövät kahekesi paramini, teeneteese õlul; Püüab teise õlul eesmärki saavutada.

VÄIT-LAUSE	**tuul on teise inimese abi**	tuul tuul+0 //_S_ sg n // on ole+0 //_V_ b // teise teine+0 //_P_ sg g // inimese inimene+0 //_S_ sg g // abi abi+0 //_S_ sg n //	täislause		
	ega jumal mõni karjapois-kene ei ole	ega ega+0 //_J_ // jumal jumal+0 //_S_ sg n // mõni mõni+0 //_P_ sg n // karjapoiskene karjapoiskene+0 //_S_ sg n // ei ei+0 //_V_ neg // ole ole+0 //_V_ o //		Kuu, Puh	Ega jummal mõni karjapoiskene ei ole, et sa teda alati hõigut; Ega jumal järvamies ole, et mihe hädasse jätab.
	ämma tarvis	ämma ämm+0 //_S_ sg p // tarvis tarvis+0 //_D_ //		Hlj, Kad, Krk	Ei oska keski üht tööd teha, peab ämma otsima.
SOOV-LAUSE	**oleks laastud lapsed olema**	oleks ole+ks //_V_ ks // laastud laast+d //_S_ pl n // lapsed laps+d //_S_ pl n // olema ole+ma //_V_ ma			

KÄSK-LAUSE	**aita ähkida**	aita aita+0 //_V_ o // ähkida ähki+da //_V_ da //		Kuu, Urv	Tule avida ähkidä; Tehi no sa, siss ma nõsta.
HÜÜD-LAUSE	**kaitse veel kõhnal karja**	kaitse kaits+0 //_V_ o // veel veel+0 //_D_ // kõhnal kõhn+l //_A_ sg ad // karja kari+0 //_S_ sg p //		Hää	

Representation of Estonian idiomatic expressions in the dictionary

Katre Õim

Summary

The main goal of this article is to describe some possibilities to compile an online dictionary of Estonian phraseology. Users of this cognitive linguistics inspired dictionary are not expected to have any knowledge on the lexical content and syntactic structure of the idiom. The dictionary is based on the electronic database of Estonian phrases.

A publicly available online dictionary brings together expressions that are semantically linked and demonstrates the morphosyntactic structure of Estonian phrases, also shows how they are used in the natural language. In the microstage the metaphor target-concept based entries would follow many distinct parameters. We hope that in the final stage the online dictionary of Estonian phraseology will represent expressions which have the same general meaning and structure.

Keywords: cognitive linguistics, lexical semantics, thesaurus, phraseology

Autor

PhD Katre Õim, Tallinna Ülikooli eesti keele ja kultuuri instituudi dotsent, Eesti kirjandusmuuseumi vanemteadur, riikliku programmi „Eesti Keele keeletehnoloogiline tugi“ projekti „Eesti fraseologismide elektroonilise alussõnastiku loomine“ vastutav täitja, katre.oim@tlu.ee