https://doi.org/10.7592/MT2017.69.eslon

Kasutuspõhise keelekäsitluse pedagoogiline perspektiiv

Pille Eslon
Tallinna Ülikooli Digitehnoloogiate Instituudi dotsent
pille.eslon@tlu.ee

Teesid: Lingvistikateadmiste sidumine erinevate keeleoskustasemete ja kooliastmetega riikliku õppekava eesti keele kui teise keele aineprogrammis on konsensuslik, see ei tugine empiirilistele uurimustele, vaid juurdunud arusaamadele ja rahvusvahelisele eeskujule. Tõeväärsemaid andmeid keeleoskustasemete kohta saab emakeelekõneleja ja õppija tekstikasutusmustrite võrdlemisel. Selleks on erinevaid võimalusi ja meetodeid. Artiklis käsitletakse eestikeelsete tekstide töötlemise suundi ja vahendeid, mis pedagoogilisest aspektist olulised, kirjeldatakse keelekasutusmustrite otsimiseks mõeldud programmi Klastrileidja ja selle abil saadud uurimistulemusi – sagedamaid keelestruktuure verbist vasakul ja paremal.

Märksõnad: eesti keele õpe, kasutuspõhine keelekäsitlus, korpuslingvistika, lingvistiline klasteranalüüs

1. Eesti keele õppe probleem

Tänapäeva keeleõpe tugineb Euroopa Nõukogu kehtestatud kuuele keeleoskustasemele, eeldades iga taseme sõnastiku, õpiku jm materjalide olemasolu, mis tagaksid järjepidevuse kommunikatiivsete, sh lingvistiliste pädevuste kujunemises neljas osaoskuses. Toetudes keeleõppe asjatundjate ühisele arusaamale ja rahvusvahelisele kogemusele, on koostatud ka eestikeelsed tasemeoskuste kirjeldused[1], välja töötatud tasemeeksamite sisu, testid, hindamise alused[2] ja nõuded hindajate töö kvaliteedile (vt Pajupuu 2007). Samas pole iga keeleoskustaseme lingvistikapädevusi süsteemselt kirjeldatud ega teaduslikult põhjendatud. Puudub ühtne arusaam, mida ning millises järjestuses omandada, kuidas jõuda niikaugele, et erineva lähtekeelega inimesed suudaksid kombineerida eesti keele vahendeid sama ratsionaalselt kui emakeelekõnelejad. Konsensuslik lingvistikapädevuste jaotamine tasemete ja kooliastmete vahel meid edasi ei aita ning selles peitub tasemeoskuste sisu ja järelikult ka hindamise nõrk lüli.

Artikli eesmärk on tutvustada mõningaid pedagoogiliste lahenduste lähtekohti ja rakenduslikult olulisi lingvistilisi uurimusi, k.a eesti kirjakeele klasteranalüüsi tulemuste perspektiivsust tasemeoskuste lingvistilise sisu teaduslikult põhjendatud modelleerimisel. Seletan Klastrileidja võimalusi tekstikasutusmustrite leidmisel ja rühmitamisel, toon näite verbi vasak- ja paremkonteksti iseloomustavatest lingvistilistest mustritest.

2. Pedagoogiliste lahenduste lähtekohti

Kirjakeele ja õppijakeele senise võrdlemise põhjal saab pedagoogilistel eesmärkidel väita, et 1) traditsioonilise normatiivse grammatika põhireegel (reeglistik) ei pruugi vastata emakeelekõneleja tekstikasutusreeglitele; 2) funktsionaalse keeleoskuse aluseks on piiratud leksikaalgrammatilise varieerumisega lingvistilised mustrid; 3) lingvistikapädevuse kujundamise raskuspunkt langeb lähte- ja sihtkeele morfofonoloogilisele sarnasusele, mitte erinemisele või põhjendamatult lihtsustatud käsitlustele.

2.1. Grammatikareeglid *vs.* keele kasutusreeglid

Grammatika tuum on morfoloogilised paradigmad. Traditsioonilistest keelekirjeldustest saab süsteemseid teadmisi vormimoodustuse ja tuletuse kohta, mitte liidete ning vormide produktiivsuse kohta tekstikasutuses. Seetõttu tulevad reeglite ja tekstiandmete võrdluses esile nii grammatika kitsaskohad kui ka kasutusreeglite sõnastamise probleemid.

Kujukas näide on soome keele *skele*-liite (mitte)produktiivsus (Lyytikäinen 2012: 114–116). Grammatikareeglite järgi pole *skele*-liide produktiivne[3], esineb tavaliselt *a*- ja *k*-alguliste kahesilbiliste verbitüvedega. Sõnastiku andmetel on *a*-algulisi tüvesid suhteliselt vähe (kokku 19, neist *skele*-liitelisi 14, ilma 5) ja *k*-algulisi tüvesid hulganisti rohkem (kokku 144, neist *skele*-liitega 99 ja liiteta 45), nt *ahmiskella*, *aioskella*, *aistiskella*, *ammuskella*; *kaadeskella*, *kaareskella*, *kaaviskella* jne. Veebimaterjalides kasutatakse *skele*-liidet mitte ainult mõlemat liiki kahesilbiliste tüvedega, vaid ka ühe- ja kolmesilbilisetega (nt *käyskellä*, *käveleskellä*). Tekstikasutuses leidub samuti *skele*-liitelisi kontraheerunud verbe[4] (nt *kiiveskellä*, *kiroskella* ja *tarjoskella*), mida rakendatakse mittekontraheerunud verbiparadigma järgi, ehkki reegel ütleb, et kontraheerunud verbid *skele*-liitele alustüveks ei sobi. Seega näitavad kirjeldava grammatika reeglid, et *skele*-liiteliste tuletiste kasutamine on morfofonoloogiliselt väga piiratud ja mitteproduktiivne ning sõnastik ja tekstikasutus, et *skele*-liiteliste tuletiste

esinemus on arvatavast tunduvalt avaram, hõlmates nii kontraheerunud verbe kui ka uudissõnu, mis viitab liite produktiivsusele. Järelikult läheb ISK 2004 grammatikareegel vastuollu soome keele kasutusreegliga. Analoogse tulemuseni on jõudnud Heiki-Jaan Kaalep (2012). Käsitlenud aktiivse (avatud) ja passiivse (suletud) morfoloogia põhimõtete rakendamise viisi eesti keele akadeemilistes väljaannetes, toob ta võrdlusena keelekasutaja jaoks loomulikud käänamisreeglid ja näitab, et need ei pruugi ühtida õigekeelsuse normidega.

Niisiis avavad empiirilise korpusanalüüsi tulemused grammatikakirjelduse kitsaskohti, mis on oluline nii teooria kui ka rakenduste mõttes. Vormistiku, paradigmade ja abstraktsete grammatikakategooriate taga pole lihtne näha keele kasutusreegleid.

2.2. Korpusanalüüs ja meetodid

Tänaseks on hulganisti uurimusi, milles põhjendatakse keelekasutusmustrite olulisust loogiliselt üles ehitatud, sidusate ja kommunikatiivselt väärtuslike tekstide produtseerimisel, sisu mõistmisel ja lugemiskiiruse arendamisel (vt nt Conklin & Schmitt 2008). Korpuspõhised sõnavara- ja grammatikauurimused on kujundanud uue arusaama sellest, kuidas inimkeel tegelikult toimib (Biber *et al.* 2006: 55–58). Tekkinud on uudsed kasutuspõhised taksonoomiad (nt Uiboaed 2013: 185; Trainis & Allkivi 2014; Eslon 2017), tekstitöötluse tulemuste põhjal on kirjeldatud lingvistilisi arenguid ja keele funktsioneerimise eripära, mida eelnevalt pole märgatud või tõestada suudetud (Küngas 2014; Küngas 2013; Valdmets 2010; Klavan 2012; Jürine 2015; Eslon & Paeoja 2015; Ruutma *et al.* 2016).

Tähelepanu all on lekseemide distributiivsed omadused ehk nende formaalne tekstiline jaotumine, mille alusel saab leida semantilisi klassifikatsioone (Šajkevitš 1976: 360). Teoreetilist väärtust omab seejuures semantika ja grammatika integreeritud käsitlemine (Greenberg 1990: 219–220). Näiteks vene keele seletav-kombinatoorses sõnastikus on iga lekseemi kirjeldatud sama skeemi alusel, mis toob esile lekseemide kombineerimise reeglid (Mel'čuk 1995: 81–133). Aktiivse sõnaraamatu koostajad analüüsivad lekseemide semantilisi seoseid, toovad esile mõiste- ning sünonüümipesasid jm (nt Apresjan *et al.* 2006). Huvi pakub tähenduste kinnistumine teatud lekseemide, vormide, funktsioonide ja konteksti tüüpidega, mis kajastavad keelemeelelist omapära ning koondavad lekseemi ümber väljendeid.

Tänased empiirilised uurimused põhinevad korpustel, jättes vaatlusandmed tagaplaanile. Siinkohal kerkib põhimõtteline küsimus (vt Widdowson 2000), kas seos teoreetilise lingvistika ja keelekasutuse vahel on otsene või vahendatud

korpuslingvistikast, mis teostab uurija individuaalsusest, konkreetsest keelest ja teooriast sõltumatut algoritmidele tuginevat analüüsi (nt Holl *et al.* 2004), sest vastasel juhul võib takerduda teooriasse (vt Tognini-Bonelli 2001: 65, 84 jj). Samas paigutub suurem osa empiirilistest uurimustest kindla teooria raamesse. Et jõuda uute tõsiselt võetavate teaduslike üldistuste ja avastusteni, läheb vaja kvalitatiivset hüpet. Mati Hint märgib, et olulisim faktor on seejuures uurija intuitsioon, kognitiivsed võimed. Olles mõtestanud teadlasele omase intuitsiooni ja deduktiivse mõtlemisviisi vahekorda induktsiooni ja tehnoloogiaga, väidab ta, et intuitsiooni järgimine viib avastusteni, "mis kuuluvad teadusse, mitte tehnoloogiasse". Keelest ja teooriast sõltumatu algoritmidele tuginev analüüs (tehnoloogia) ei vii teaduslike avastusteni, mida võib saavutada uurija tänu deduktiivsele mõtlemisviisile. Tehnoloogia rakendamisega kaasneb leiutamine. (Hint 2016: 632.)

Kui aga loobuda teooriat tõendavatest / ümber lükkavatest andmetest ja rakendada tehnoloogiat, siis võib mahukate korpuste põhjal leida keeleelementide loomulikud kombineerimisvõimalused. See aines vajab mõtestamist kasutusgrammatika ehk aktiivse grammatika võtmes. Tegu on induktiivse suunaga, kus korpusanalüüsi empiirilised tulemused näitavad, miks me kasutame keelt just nii, nagu me kasutame. Selle põhjal saab teha järeldusi nii aktiivse grammatika kui ka keele kasutusstandardi kohta, millel heuristiline tähendus ka teoreetilisele lingvistikale – mitte teooria tõestuse või ümberlükkamise eesmärgil, vaid avastamise mõttes. Pedagoogilises plaanis annab see lähenemisviis lingvistilist alusmaterjali funktsionaalse keeleoskuse kujundamiseks.

Lingvistika ja tehnoloogia piirimail on kujunenud korpuslingvistiline avastusele orienteeritud suund, mille meetodid ja tehnoloogia on universaalsed. Neid saab kasutada erinevatel eesmärkidel ja erinevates valdkondades – nii humanitaar- (nt allkeeled ja indiviidi keelekasutus, keeleõpe, õpikutekstide keerukus, tõlkimine, kultuurierinevused) kui ka sotsiaalvaldkonna uuringutes (nt poliitiline diskursus, sündmused ja tegijad, haiguslood) ja mujalgi. Avastuspõhistes (ingl *detection-based approach*) esimese ja teise keele uurimustes on analüüsitud kirjalike ja suuliste tekstide leksikaalgrammatilist rikkust, lingvistilist keerukust, terviklikkust ja ülesehituse loogikat, esituslaadi korrektsust ja sujuvust, eri lähtekeelega õppijate keeleoskuse edenemist, veamustreid, võrreldud leksikaalgrammatilist varieerumist jm. Aluseks on ICLE-korpuse[5] eri lähtekeelega inglise keele õppijate kirjalik ja suuline keelekasutus, mille analüüsimiseks rakendatakse erinevaid andmekaeve tehnikaid ja statistilisi meetodeid (tavaliselt lineaarset diskriminant-, korrelatsiooni- ja regressioonianalüüsi). (Vt Jarvis & Crossley 2012.) Analüüsi suuna ja tehnikate valiku tingivad uurimistöö objekt ning eesmärk, milleks mõned on efektiivsemad,

kuna annavad usaldusväärsemaid tulemusi (vt Aedmaa 2015: 42–52; Aedmaa 2016: 12–13; Klavan 2012; Ruutma *et al.* 2016: 97–108).

Samas ütlevad skeptikud, et loomuliku keele tekstide töötlemine tarkvaraga ei taga üheseid tulemusi ja statistilised andmed võivad olla rohkem või vähem tõeväärsed. Järelikult ei tea me ikkagi täpselt, kuidas hakkab inimene konkreetsetel suhtlusvajadustel keelevahendeid kombineerima. Tekstiloome tugineb vaid osaliselt kasutusstandardile, mustritele ja registrimarkeritele, ülejäänus on see protsess individuaalne ning erinevatest faktoritest mõjutatud. Seetõttu on perioodi väljendatud põhjendatud kahtlust, kuivõrd ikka sobivad korpused ja korpuslingvistilised meetodid teoreetilise lingvistika ülesannete lahendamiseks (nt Chomsky 1957: 15 jj; Hint 2016: 632 jj).

Loomulikult pole olemas ühte-kahte universaalset meetodit, mis tooksid formaliseeritud kujul esile kõik inimkeele kasutusreeglid, ega korpusi, milles nende reeglite toimimine leiaks sajaprotsendilist tõestust. Ometi on saanud korpuslingvistikast iseseisev suund, mille raames arvutilingvistilisi ja statistilisi meetodeid kombineerides lahendatakse keeleõppe, automaattõlke, masinõppe jt probleeme, kaasa arvatud lingvistilisi, nagu vormihomonüümia, defektsed paradigmad, määramata juhtumid jne, mille kohta saab uusi andmeid korpustest, nt eesti verbi lihtvormide paradigma (vt Kaalep 2015) või eesti keele ambipositsioonide *läbi*, *mööda*, *vastu*, *üle* ja *ümber* funktsionaalne, semantiline ning morfosüntaktiline piiritlemine klassifitseerimispuude meetodil (vt Ruutma *et al.* 2016).

Analoogselt võib iga keeleoskustaseme lingvistilisi mustreid käsitleda võrdluses kirjakeelega, hinnata nende olulisust tekstiloomes, seletada morfosüntaktilise varieerumise/kinnistumise vahekorda, mustri komponentide semantilist sidusust jm, mis on tasemete lingvistilise modelleerimise üks aspekte.

2.3. Raskustest lingvistikapädevuse kujundamisel

Lingvistikapädevuste kujundamist komplitseerivad selle erinevad tahud. Nimetan mõned neist.

2.3.1. Lähte- ja sihtkeele vahelised seosed

Õpiraskused tulenevad suuresti lähte- ja sihtkeele vaheliste morfofonoloogiliste seoste iseloomust (sümmeetria-, asümmeetria- ja analoogiaseosed), ent metoodilises kirjanduses pole konkreetselt sellele küsimusele tähelepanu juhitud, ehkki keeltevahelise mõju uurimine laieneb (nt Bartning *et al.* 2010; Jarvis & Crossley 2012; Kaivapalu 2013; Kaivapalu & Martin 2014). Ka pole eesti

keelt võrreldud lähiümbruse kontaktkeeltega ei loomuliku morfoloogia ega markeerituse teooria seisukohalt, v.a soome keel (Remes 2009). Tänu Hannu Remesi uurimusele saab konkreetseid lingvistikateadmisi, mis seletavad, et lähte- ja sihtkeele sümmeetrilisus võib praktikas tekitada suuremaid õpiraskusi kui asümmeetrilisus. Näiteks keeltevahelise markeerimatuse/markeerituse aspektist on soome ja eesti keele ***i***- ning ***si***-imperfekti vormide kasutamine vastupidine. Soome keeles on tavaline ehk markeerimata ***i***-tunnus (*anta- : anto/i*, *elä- : el/i*, *huuta- : huus/i*) ja eesti keeles jälle ***si***-tunnus (*ela/si/n*, *luge/si/n*, *kirjuta/si/n*); ***i***-tunnus esineb väheste, kuid tekstikasutuses väga sagedate verbidega (*saa-: sa/i/n*, *ole-: ol/i/n*, *pane-: pan/i/n*). Seetõttu kaldub eestlane emakeele mõjul ka soome keeles eelistama ***si***-imperfekti ja soomlane eesti keeles ***i***-imperfekti. Nende vormide valik tekitab õppijale lisaraskusi vormimoodustuse tõttu, sest nii soome ***si***-kui ka eesti ***i***-imperfektiga kaasneb laadivaheldus. (Remes 2009: 41.)

Näivat sarnasust esineb ka mittesugulaskeelte vahel, kuid veelgi keerulisemad on juhtumid, kus lähte- ja sihtkeele vahelisi seoseid pedagoogilistel eesmärkidel meelevaldselt interpreteeritakse. Näiteks samastades vene keele eessõnade ja eesti keele kaassõnade kasutamist 1) prepositsioonis (***с*** *мам**ой*** ~ *ema**ga***, ***koos*** *ema**ga***; ***без*** *отц**а*** ~ *isa**ta***, ***ilma*** *isa**ta***) ning 2) postpositsioonis (***под*** *стол**ом*** ~ *lau**a all***; ***под*** *стол* (***ø***-lõpp) ~ *lau**a alla***; ***из-под*** *стол**а*** ~ *lau**a alt***). Keeltevahelise markeerimatuse/markeerituse seisukohalt vastandub eesti markeeritud prepositsioonifraasile vene keele markeerimata prepositsioonifraas. Seda asümmeetriaseost on võimalik pedagoogilistel eesmärkidel kasutada, kui aluseks võtta fraasistruktuuri analoogia – vene eessõna / eesti kaassõna + nimisõna. Ent samamoodi pole võimalik käsitleda eesti keele markeerimata postpositsioonifraasi, sest eesti ja vene keeles on siin tegu erinevate fraasistruktuuridega ja analoogia ei toimi.

2.3.2. Normeeritud kirjakeele standard vs. *kasutusstandard*

Pedagoogilisel eesmärgil orienteerutakse õigustatult normeeritud kirjakeele standardile, ent põhjendatum oleks kasutusstandard, sest norm on kokkuleppeline, kasutusmustrid aga ühiskondlikus praktikas aja jooksul välja kujunenud ja kinnistunud. Kahjuks pole kasutusstandardit varnast võtta, kuid igaüks võib rakendada oma tehnoloogiateadmisi ja arvutioskust ning leida vajalikku materjali eesti keeleressurssidest. Siinkohal tuleb aga tõdeda, et kuigi kooli- ja ülikooliharidus sisaldab arvutiõpet, digipädevusi kinnistatakse erialase ettevalmistusega ning kuigi meil on keelekorpused[6], sagedussõnastik[7], fraseologismide elektrooniline alussõnastik[8], püsiühendite andmebaas[9], kaks

uut sõnamoodustusraamatut[10] jm, juurdub keeleressursside kaasamine eesti keele õppesse üsna visalt.

Keeleveebi e-sõnastikke kasutades on vaja lingvistikateadmisi, sest märksõnaseletused sisaldavad otse- ja ülekantud tähendusi, viidet muuteparadigmale, erinevat grammatilist infot, ka jooniseid jm illustreerivat teavet. Raskusi passiivset tüüpi sõnastike kasutamisel on kirjeldanud Katre Õim (2009), pakkudes välja alternatiivseid mooduseid fraseologismide esitamiseks. Samuti on teada (vt Apresjan *et al.* 2010: 17–18), et passiivse sõnastiku 100 000 märksõnast moodustavad 60–70% tuletised, murdesõnad ja terminid, mida me igapäevaselt ei kasuta. Keeleõppe eesmärgil läheb tarvis kõnearendusele ja tekstiloomele suunatud aktiivset üldkeele sõnastikku, mis sisaldaks võimalikult palju teavet kasutuspiirangute kohta, nagu nt Longman (2003)[11] või vene keele seletav-kombinatoorne sõnastik[12]. Eesti keele põhisõnavara sõnastik (2014)[13] on samm teel aktiivse sõnastiku koostamiseks (5000 märksõna, mõeldud A2–B1 tasemele). Siit saab abi nii rääkimisel (sisaldab häältdusnäiteid) kui ka kirjutamisel (sisaldab verbi sagedamaid argumendistruktuure koos lihtsate kasutusnäidetega). Sõnastik kajastab üld- ja põhisõnavara tekstikasutust ning on selles mõttes aktiivset tüüpi, kuid puudub jaotus A2- ja B1-taseme vahel. Et niikaugele jõuda, tuleks alustada üld- ja põhisõnavara eristamisest. Iseküsimus muidugi, kuidas siduda pedagoogilisel eesmärgil eri tasemete üld-, põhi- ja temaatiline sõnavara ning õppekavaga määratud läbivad teemad.

Eesti keele tasemekirjeldustes leidub teemavalikute all ka keelenäiteid, A2-tasemele on lisatud isegi esmase või elementaarsõnavara loend[14], kuid jääb selgusetuks elementaarsõnavara mõiste, eristamispõhimõtted ja maht. Kirjeldatakse, milliste allikate põhjal on loend koostatud (eesti kirjakeele sagedussõnastik (2002)[15], "Esmane verstapost. Eesti keele suhtluse algtase" (1998)[16]), ligi 500 sõna on autor lisanud oma õpetamiskogemuse põhjal. Varematel aegadel on ilmunud leksikograafide töökogemusele rajanev eesti keele baassõnastik[17] ehk põhisõnavara sõnastik, kuid selle maht tekitas küsimusi (nt kas piisab 1500 või 3000 sõnast). Samas arvutasid soome keele sagedussõnastiku koostajad kohe välja, et 90% soomekeelse teksti mõistmiseks piisab 12 663 sõnast (vt Saukkonen *et al.* 1979: 7), mida hakati kasutama soome keele õppes. Analoogselt koostati vene keele sagedussõnastiku (1965) alusel koolisõnastik (1965), milles sisalduv aktiivne sõnavara on jaotatud kooliastmeti. Sellele sõnavara miinimumile rajanesid nii aineõpetus kui ka vene keele õpikute koostamine[18].

Lahendusi üld- ja põhisõnavara eristamiseks on pakkunud Kairit Sirts ja Leo Võhandu, kuid see pole leidnud järgimist. Autorid seostavad põhisõnavara semantiliste primitiividega, mis on üldsõnavara tuum. Iga allkeel kujuneb selle põhjal. (Sirts & Võhandu 2009: 261–262.) Siin leidub analoogiat Igor Mel'čuki ja Aleksandr Zholkovsky seletav-kombinatoorse sõnastiku (1984) formalismi

ning standardsuse põhimõtetega, mis ei jäta sõnaseletuses ruumi oletustele. Märksõnaks valitud ühetähenduslikud ja nendega tähenduse poolest sarnased lekseemid ühenduvad mõistepesadeks. Lekseemi semantilisele kirjeldusele järgneb süntaktiliste ja leksikaalsete piirangute ehk kombineerimisreeglite kirjeldus (Mel'čuk 1995: 5–6). Seega annab aktiivne sõnastik kasutusstandardi, mis aluseks tekstiloomele.

2.3.3. Õpikud

Inimene kombineerib ning varieerib erinevaid sisu- ja funktsioonisõnu suhtlusvajadustest tulenevalt. Osa neist sõnadest kuulub registrimarkerite alla, osa on konventsionaalsed vormelid (vt Conklin & Schmitt 2008: 73–76, 86), osa rohkem või vähem iseloomulikud stereotüübid, mille põhjal saab moodustada uusi mittekonventsionaalseid järjendeid nagu analüütilised sõnad, vormid ja konstruktsioonid (Eslon 2014a: 17–18; Eslon 2017: 45–58). Need üksused on suunatud nii tekstidest aru saamisele kui ka tekstiloomele, kuid ei lahenda küsimust, millistel põhimõtetel piiritleda temaatilist sõnavara üld- ja põhisõnavarast. Veelgi enam: tegelikult kasutavad eesti keele kui teise keele õpetajad tänaseni põhikooliõpikuid, mille tekstide korpuslingvistiline analüüs tõi esile kohati olematu seose õppekavas ette antud teemade ja temaatilise sõnavara laiendamisega, üld- ja põhisõnavara kordamise ning kinnistamisega, grammatikaõpetusest rääkimata (vt Ševtšenko 2014; Tšernõšuk 2016). Samas on kindlaks tehtud, et õppija sõnastused on suurel määral identsed õppetekstide sõnastustega, nt rootsi lähtekeelega algtaseme soome keele õppijate tekstiloomes 34% ulatuses (Määttä 2012: 213).

Meie olukorda on võimalik muuta, kui koostada kontseptuaalselt sidusad iga keeleoskustaseme omandamisele suunatud õpikud ja õppesõnastikud. Enne peaks aga 1) kindlaks tegema iga taseme lingvistilise sisu (keelekasutusmustrite leksikaalsed, morfosüntaktilised ja funktsionaalsed piirangud); 2) eristama aktiivse sõnavara ja aktiivse grammatika, sidudes need teemade läbimiseks vajalike tekstidega jne. Eesti keele õpetamine ei peaks 21. sajandil tuginema kollektiivsele õpetamiskogemusele, individuaalsetele tõekspidamistele ja neil põhinevatele interaktiivsetele õpikeskkondadele, samuti eksitavatele tõlgendustele, vaid teaduspõhise analüüsi tulemuste pedagoogilisele rakendusele.

3. Pedagoogilise rakendusega eesti keele empiirilised uurimused

Empiirilistest uurimustest on pedagoogilisel eesmärgil kasu eelkõige nendest, mille väljund on õppeleksikograafia, kasutusgrammatika, kasutusstandard, tekstide sisu mõistmine ja tekstiloome, keeleoskustasemete ennustamine, kontrastiivanalüüs.

1) Eesti keele õppeleksikograafilise käsitluse aluseks olevas Jelena Kallase (2013) uurimuses on korpustest otsitud tüüpilisi keelekasutusmalle (vt Kallas *et al.* 2012), mis näitavad sõnade semantilist sobivust, verbide argumendistruktuure ja olulisi grammatilisi seoseid. See materjal (vt Kallas *et al.* 2014) võimaldab lisaks sõnavarale kujundada ka üldist grammatikapädevust (vt Kallas *et al.* 2015). Õppesõnastike koostamisel läheb vaja spetsiaalseid keeleressursse nagu fraseologismide elektrooniline alussõnastik ja püsiühendite andmebaas, fraseologismide ja püsiühendite otsing semantiliste rühmade kaupa jm (vt Kaalep & Muischnek 2009; Õim 2014). Koostatakse kollokatsiooni- (Kallas *et al.* 2015; Koppel & Kallas 2016; Kallas, J. *et al.* 2017) ja assotsiatsioonisõnastikku (Vainik 2017), ilmunud on emotsioonisõnavararaamat (Vainik 2016).

2) Rakenduslikul, sh pedagoogilisel eesmärgil on kasu kirjakeele ja õppijakeele kasutusmustrite võrdlemisest ja lingvistilisest tõlgendamisest, mis seletab, kui sarnane, nihkes või erinev on keele elementide valik ning kombineerimise viisid (nt Eslon 2013, 2014a, 2014b; Allkivi 2016a, 2016b). Senised uurimistulemused on esile toonud keelekasutuses olulisi nähtusi, nt adverbi sisaldavate struktuuride juhtiva rolli emakeelekõneleja tekstiloomes, adverbide funktsionaalse potentsiaali loogiliselt üles ehitatud sidusate tekstide koostamisel, vabadest sõnaühenditest liitsete analüütiliste üksuste moodustumisel jm (nt Eslon 2014a, 2014b, 2017; Paeoja 2015; Eslon & Paeoja 2015; Trainis 2015; Trainis & Allkivi 2014).

3) Järgmine rakendusliku väljundiga uurimissuund on keeleoskustaseme ennustamine masinõppe meetodeid kasutades. Seda suunda on õppijakeele analüüsimisel praktiseeritud valdavalt inglise keele põhjal (vt eespool), praeguseks on lisandunud saksa, rootsi ja eesti keel. Eesmärk on tasemeoskuste automaatne testimine (vt Vajjala & Lõo 2014: 114–115), mis võimaldab eesti keele õpet ja hindamist individualiseerida. Eesti vahekeele korpuse (EVKK) tuumkorpuse A2-, B1-, B2- ja C1-taseme tekstide alusel on analüüsitud õppija sõnavara rikkust, leksikaalset mitmekesisust, süntaktilist keerukust, veamustreid jm. Selleks on Sowmya Vajjala ja Kaidi Lõo mõõtnud eesti õppijakeele morfoloogiliste seoste tugevust ning esile toonud tasemeid eristavaid morfoloogilisi tun-

nuseid. Varem on Vajjala ja Lõo (2013) analüüsinud eesti õppijakeele sõnavara mitmekesisust ning grammatikastruktuuride keerukust EVKK tuumkorpuse kolmeastmelise A, B ja C skaala alusel.

4) Et õpetada masinat määrama keeleoskuse taset, leidma vigu ja neid parandama, peab masin suutma ühestada mittestandardset sõnakasutust. Selleks läheb vaja õigekirjakorrektorit, mille kallal keeletehnoloogid töötavad (nt Liin 2009). Kairit Sirts (2012) on katsetanud robustset lemmatiseerijat. Prototüüp põhineb foneetilisel algoritmil Metaphone (välja töötatud inglise keele alusel, eesti keele jaoks lisatud täishäälikud *õ*, *ä*, *ö* ja *ü*). Põhimõte seisneb selles, et taandada analüüsitav sõna standardkeele häälikulisele kujule, kusjuures sarnaselt kõlavatel sõnadel on ühesugune häälduskuju. Nii rühmituvad sarnase häältusega sõnad hulkadeks. Õppijakeele valesti kirjutatud või moodustatud vormidele leitakse häälduskuju võrdluses standardkeelega. Tõenäolist kandidaati otsitakse sõnade hulgast, mille teisenduskaugus standardkeele sõnadest on maksimaalselt 2, s.t võrdne operatsioonide arvuga, mis vajalikud ühe stringi teisendamiseks. Arvestatud on tähe lisamise, kustutamise ja asendamisega (Levenshtein 1966). Näiteks: *kantsid* vs. *kandsid* _**t** asendada **d**-ga _kaugus = 1; *igasugulased* vs. *igasugused* _kustutada l ja a _kaugus = 2.

Niisiis kasutavad esimesed kaks empiirilist uurimissuunda korpusi, keeletarkvara ja korpuslingvistilisi meetodeid, kuid erinevaid vahendeid – Sketch Engine'i ja Klastrileidjat. Kolmas uurimissuund lähtub õppijakeele morfosüntaktilisest ja semantilisest arvutianalüüsist, mida praktiseeritakse Tübingeni Ülikoolis individuaalse keeleõppe- ja automaatse testimisrakenduse loomiseks (Detmar Meurers jt). Robustse lemmatiseerijaga seotud arendused on kujunenud koostöös Eesti ning Austraalia keeletehnoloogidega. Kirjaliku ja suulise keelekasutuse uurijatele pakutakse järjest täiuslikumat tehnoloogiat ning vahendeid: reeglipõhine ja statistiline tekstianalüüs, sõltuvuspuude pank, TEXTA, EstNLTK, leksikograafi tööriistad, masintõlge, kõnetuvastus ja -süntees. Järgnevalt tutvustan lühidalt Klastrileidja tööpõhimõtteid.

4. Klastrileidja

Klastrileidja[19] on programm, mis töötab andmekaeve põhimõttel, ühendades klastriteks ja reastades sageduse järgi kõik samalaadse lingvistilise märgendusega n-grammid. Sellest ka programmi nimetus.

Klastrileidja on Erika Matsaku sõnajärjeleidja prototüübi (Metslang & Matsak 2010; Matsak *et al.* 2010) arendus, töötab nii Java programmina kui ka EVKK veebirakendusena (vt Ots 2011, 2012). Esimese sisendiks on reeglipõhise

EstCG 1,0 parseriga pindsüntaktiliselt eelmärgendatud tekstid, teine suudab lugeda sagedusega kombineeritud varianti.

Klastrileidja otsib morfosüntaktiliselt märgendatud tekstist sarnaseid morfo- ja süntaksimärgendite järjendeid, ühendab need järgemööda, fikseerib sageduse ja lisab keelenäited. Enne märgendatud tekstide sisestamist tuleb valida n-grammi pikkus (bigramm, trigramm, tetragramm jne), analüüsi lingvistiline objekt (morfoloogia, süntaks, morfosüntaks) ning läbi mõelda, kas on vaja arvestada ka (osa)lause piiriga.

Lingvistilise objekti valik ning n-grammi pikkus (tavaliselt bi- ja trigrammid) olenevad uurimise eesmärgist. Toon selle kohta väikese näite tekstikasutusmustritest, mis leitud Java programmina töötava Klastrileidjaga EVKK vene emakeelega gümnaasiumiõpilaste eesti keele olümpiaaditööde alamkorpusest.

Kui objekt on süntaktilised funktsioonid, siis valitakse otsing süntaksimärgendite alusel: **CLB @J @SUBJ @+FMV (451 näidet, nt *et autor tahab*); @ SUBJ @+FMV @ADVL (446 näidet, nt *autor kirjeldab mitte*); @+FMV @ADVL @ADVL (415 näidet, nt *on tänapäeval nii*).

Kui objekt on morfoloogilised struktuurid, siis kasutatakse klasterdamist sõnaliigimärgendite alusel: verb (V) + adverb (D) + adverb (D) ehk VDD-struktuur (64 näidet, nt *on veel vara*); eitus (V) + verb (V) + adverb ehk VVD-struktuur (57, nt *ei tule enam*); adverb (D) + adverb (D) + adverb (D) ehk DDD-struktuur (52, nt *juba kusagilt mujalt*).

Kui objekt on vormid, siis piirdub otsing morfoloogiliste märgenditega: _V_ aux neg + _V_ main indic pres ps neg #FinV #Intr + _D_ (*ei tule enam*), st VVD-struktuuri vormikasutust iseloomustab eitav kõne, intransitiivne verb indikatiivi preesensis ja adverbiaalne laiend.

Kui objekt on morfosüntaks, siis kasutatakse klasterdamist morfo- ja süntaksimärgendite alusel: _V_ aux neg @NEG + _V_ main indic pres ps neg #FinV #Intr @+FMV + _D_ @ADVL (nt *ei tule enam*), s.t eitav kõne, intransitiivse verbi indikatiivi preesens lihtöeldisena, järgnev adverb on adverbiaali funktsioonis. Kvalitatiivse analüüsi käigus selgub määruse liik (ajamäärus).

Kuna Klastrileidja otsib ühesuguse lingvistilise märgendusega n-gramme, ühendab need esinemissageduse põhjal ja lisab keelenäited, siis on tegu *lingvistilise* klasteranalüüsiga (vt Allkivi *et al.* 2017; Trainis käesolevas kogumikus). Klastrites sisalduvate n-grammide hulk on arvuliselt fikseeritud, neid saab kirjeldada igal lingvistilise analüüsi tasandil, leida nende leksikaalsemantilise ja grammatilise varieerumise piirid, sõna- ja vormivaliku piirangud ning kinnistumise juhtumid. Selle põhjal on näha, mille poolest nt keeleoskustasemete, üksikisikute või allkeelte kasutusmustrid sarnanevad ja erinevad. Andmete tõesust kontrollitakse statistiliste mõõtmistega. Klastrileidja toob esile olulised lingvistilised mustrid, mille põhjal on võimalik sõnastada eesti keele kasutus-

reegleid. Keeleõppe eesmärgil on mõttekas tugineda just nendele ja õppida ennast väljendama emakeelekõnelejale omaselt.

Vastavalt analüsaatori tööpõhimõttele tuleb esile n-grammide, klastrite, alamklasside ja klasside hierarhia (alt üles klassifikatsioon), mis pole väljamõeldis, vaid põhineb loomuliku keelekasutuse seaduspärasustel, mis Klastrileidja on välja otsinud. Seetõttu on n-grammide, klastrite, alamklasside ja klasside hierarhia objektiivne ontoloogiline alus taksonoomiatele, millele saab järgnevalt üles ehitada kogu kasutusgrammatilise keelekäsitluse. Lingvistilisel klasteranalüüsil põhinevates eesti keelekasutuse kirjeldustes on alt üles klassifikatsiooni rakendatud vastupidi: klassid, alamklassid, klastrid ja n-grammid.

Klastrileidja veebirakendust võib soovitada õpikute ja õppematerjalide koostajatele, kelle ülesanne on siduda süsteemseks tervikuks temaatilised tekstid, nende sõnavara ja olulised morfosüntaktilised struktuurid, mida teemast rääkides või kirjutades vaja. Pole tarvidust nuputada, kuidas liigendada sõnavara ja grammatikat ning millises järjestuses need õpikus esitada. Õpetajal on võimalus suunata õpilasi otsima käsitletava teemaga seotud olulisi sõnastusi, võrdlema neid enda kirjutatud või valitud teksti(de) sõnastustega jm. Sageli esinevate morfoloogiliste struktuuride alusel saab harjutada leksikaalset ja grammatilist varieerumist, muuta sõnajärge, transformeerida keelestruktuure jm, et näha, kas asendused on ka tegelikult võimalikud, kas see, mis õpikus kirjas ning milliseid valikuid on õppija teinud, vastab emakeelekõneleja valikutele. Teatud määral kompenseeriks see kontseptuaalselt sidusate eesti keele õpikute ja sõnastike puudumist, kuid ei lahendaks veel eesti keele õppe probleeme.

5. Näide eesti kirjakeele kasutusmustritest verbist vasakul ja paremal

Eesti keele tekstikasutuse kirjeldamist alustasin aastate eest, kui olin põhitäitja Katre Õimu ETF-i grandis 8222 “Ülekantud tähenduses fraasid eesti keele korpustes” (2010–2013). Klastrileidja loomine inspireeris jätkama. Olen analüüsinud verbilõpulist ja verbialgulist morfoloogilist klassi. Kuna mõlemas positsioonis toob mustrite kirjeldus esile adverbi sisaldavate ja adverbita struktuuride vastanduse, siis on see formaalne tunnus mustrite liigendamise alus.

5.1. Alamklassid ja klastrid verbist vasakul/paremal

1) Verbilõpulise klassi trigrammid jagunevad kuue alamklassi vahel: D-V (44%), J-V (28%), S-V (17%), P-V (5%), A-V (5%) ja V-V (1%). Keelekasutus-

tendentside kirjeldamiseks piisab alamklassidest, mille olulisuskoefitsient k ≥ 76[20] – seetõttu jäävad välja kolm vähese esinemusega alamklassi (pronoomeni-, adjektiivi- ja verbialguline). Adverbi-, konjunktsiooni- ja substantiivialgulise alamklassi statistiliselt olulisemaid adverbe sisaldavaid klastreid on seitse (DDV, DVV, JDV, DJV, SDV, DSV ja DAV) ning adverbita kaks (JPV, JSV). Kõikide adverbi sisaldavate ja adverbita klastrite omavaheline suhe verbist vasakul on 62% *vs*. 38%.

2) Verbialgulise klassi alamklasside olulisuskoefitsient k ≥ 64[21]. Sellele tingimusele vastavaid alamklasse on seitse: V-D (59%), V-S (21%), V-A (12%), V-K (4%), V-V (2%), V-P (1%), V-J (1%). Kolme suurema esinemusega alamklassi klastrite esinemus on erinev. Adverbi-, substantiivi- ja adjektiivilõpulise alamklassi statistiliselt olulisemaid adverbe sisaldavaid klastreid on kuus (VDD, VVD, VDA, VSD, VPD, VDS), sama palju leidus ka adverbita klastreid (VAS, VPS, VAJ, VSS, VVS, sh VSK adpositsioonilõpulisest alamklassist). Kõikide adverbi sisaldavate ja adverbita klastrite omavaheline suhe verbist paremal on 81% *vs*. 19%. Seega on adverbi sisaldavate ja adverbita mustrite vastandus verbist paremal veelgi reljeefsemalt esile tulnud kui verbi vasakkontekstis.

5.2. Järeldusi alamklasside ja klastrite esinemusest verbist vasakul/paremal

1) Kui võrrelda adverbiga ja adverbita struktuuride osakaalu verbist paremal ning vasakul, siis on mõlemal juhul ülekaalus adverbi sisaldavad mustrid (vastavalt 81% ja 62%); adverbita mustrid jäävad neile kõvasti alla (vastavalt ja 19% ja 38%). Niisiis on adverbi olemasolu/puudumine verbi lähikontekstis eesti keelele omane tunnusjoon, mida saab kasutada mustrite klassifitseerimisel.

2) Mõlemas positsioonis tuleb esile klastrite sünkroonsust, mis kajastub osaliselt ka nende kasutatavuses: DDV (23%) – VDD (27%); DVV (12%) – VVD (15%); DSV (2%) – VSD (11%); SVV (4%) – VVS (1%); SSV (2%) – VSS (1%). Järelikult on verbi lähikonteksti olulisemad mustrid DD-V-DD (kaks järjestikust adverbi raamistavad verbi); DV-V-VD (adverb raamistab liitpredikaati või liitseid verbivorme); DS-V-SD (adverb ja substantiiv raamistavad verbi finiitvormi); SV-V-VS (substantiiv raamistab liitpredikaati); SS-V-SS (kaks järjestikust substantiivi määrusliku täiendi ja genitiivatribuudi funktsioonis raamistavad verbi).

3) Verbi vasak- ja paremkonteksti ilmestavad adverbi sisaldavad mustrid, neist suurim esinemus on DDV- ja VDD-struktuuril. Adverb on verbi lähikontekstis kasutatuim sõnaliik, millel avar funktsionaalne potentsiaal.

5.3. Valik mustreid verbist vasakul

Kuna kõige iseloomulikum mustrite eristamise tunnus on adverbi sisaldavate ja adverbita struktuuride vastandus, siis peavad sellel olema omad põhjused ja seletused, mille kohta annab täpsemat lingvistilist teavet *n*-grammide tasandi morfosüntaks (vt Eslon 2017, 2014a, 2014b, 2013).

5.3.1. Näiteid adverbi sisaldavatest mustritest

Olulisimad adverbi sisaldavad mustrid verbist vasakul kuuluvad alamklassi D-V (adverb-verb). Struktuuri keskmise komponendina varieeruvad tavaliselt adverb (DDV) ja verb (DVV). Alamklassis leidub ka tekstikasutuses vähem olulisi mustreid, mis jäävad praegu välja. Selle asemel toon näiteid alamklasside J-V ja S-V olulisematest mustritest JDV ja SDV.

1) Eesti keele õppija seisukohalt on DDV kahtlemata keeruline juhtum, sest nii komponentide leksikaalsemantiline kui ka morfosüntaktiline varieerumine on lai ning mustri esinemus verbist vasakul kõige suurem. Järelikult on selle kasutamist raske vältida.

Verbi, eriti analüütiliste verbide leksikaalne varieerumine on rikkalik, enamik neist ainukordsed ühendverbid (*ette tulema* (*teatama*), *edasi minema* (*tormama*), *järele kuulama* (*hõikama*), *kinni mähkima, järele pärima*). Ka adverbide leksikaalne varieerumine on avar, kuid trigrammi alguses korduvad rõhusõnad *enam*, *nüüd* ja partikkel *mitte*, keskel tavaliselt *lt*-lõpulised adverbid, nagu *põhjalikult*, *täpselt*, *õigeaegselt*, viisi- ja ajamääruse funktsioonis. Adverbid paiknevad abiverbi ja mineviku liitaja vormi kuuluva *nud*-partitsiibi vahel (*<on> kuskile ära sõitnud, <olid> seal varemgi olnud*).

Verbi morfosüntaktiline varieerumine on päris keeruline. Kõige sagedamini kasutatakse trigrammi lõpus *nud*- ja *tud*-partitsiipi, mis kuuluvad aktiivi ja passiivi mineviku liitajavormi (*<olid> seal varemgi olnud*, *<on> põhjalikult järele päritud*); *da*-infinitiiv esineb kas infiniitse öeldise (*kuhugi ära peita*), objekti (*<tahab> veelgi üle lugeda*) või subjekti funktsioonis (*<vaja> ajutiselt sisse seada*); *ma*-infinitiiv infiniitse öeldisena (*veel täna minema*). Finiitverbi ajavormi valik piirdub imperfekti ja preesensi 3. pöördega (*Ju siis puudus*; *üha enam kaugeneb*).

2) DVV on pigem eitava kõne muster, mille tekstikasutus näitab, et mida suurem on verbi leksikaalsemantiline varieerumine, seda suurem on kasutatud adverbide hulk ja vastupidi. Seetõttu on DVV alusel mõttekas õppida verbe ja adverbe kombineerima ning oma sõnavara laiendama. Verbi morfosüntaktiline

varieerumine on avar: eitava kõne indikatiivi imperfekt (*veel ei liikunud*, *välja ei koorunud*) ja preesens (*enam ei ütle*, *sisse ei saa*), harva konditsionaali preesens (*Nüüd ei oleks*, *alt ei veaks*); vähestes jaatava kõne näidetes pluskvamperfekti liitajavorm (*sealt oli põgenenud*, *Nii oli kestnud*). Silma hakkab sünkroonsus üldolevikulist preesensit sisaldava DVV-struktuuri piiratud leksikaalse varieerumise ning kasutatud verbide ja adverbide suure esinemuse vahel. Need tunnused viitavad mustri kinnistumisele üldistavas esituslaadis.

3) JDV-struktuur esineb nii jaatavas kui ka eitavas kõnes. JDV alusel on samuti lihtne sõnavara laiendada, sest üldjoontes iseloomustab mustrit üsna rikkalik adverbide leksikaalsemantiline varieerumine. Erandid: ajamääruse funktsioonis kinnistunud proadverb *siis* ning sageli kasutatud subjektiivmodaalsed tõeväärtushinnangusõnad *ilmselt*, *kahtlemata*, *kindlasti* koos verbiga *olema* indikatiivi imperfekti ainsuse 3. pöördes, harva preesensi ainsuse 3. pöördes (*kuid äkki on*) ja imperfekti mitmuse 3. pöördes (*ja siis oli*). Selle põhjal võib öelda, et tegu on morfosüntaktiliselt vähevarieeruva mustriga, mille muudavad keerulisemaks järgmised asjaolud.

a) Kuigi eitavas kõnes on kõik JDV komponendid muuteparadigmata (*ent tookord ei <tulnud, olnud; tehtud>*), kuulub eituspartikkel aktiivi/passiivi mineviku liitaegade koosseisu, mille moodustamist ja kasutamist pole kerge omandada (vrd *ent tookord ei tulnud* mida teha?; *ent tookord ei tuldud* + adverb, noomeni- või verbivorm).

b) JDV pole eitavas kõnes kuigi levinud.

c) Hakkab silma, et (osa)lause alguses moodustab rinnastav sidesõna koos järgneva adverbiga funktsionaalse ja häälduslikuterviku (*ent tookord*, *kuid äkki*, *Aga ilmselt*), mida kasutatakse viiteseosena eelnevalt kõneks olnud asjaolude täpsustamisel (*ja kindlasti oli*, *ning siis kadus*, *Kuid kahtlemata oli*).

4) SDV on eespool kirjeldatud DDV-st leksikaalselt varieeruvam ja morfosüntaktiliselt veelgi keerulisem, kuid sel on kindlad kasutusreeglid: a) subjekt ainsuse nominatiivis + sünteetiline verb; b) totaalobjekt ainsuse genitiivis või käändeline määrus ainsuse komitatiivis + analüütiline verb. Keerukust lisab mustri sõnajärg, kuna subjekti/objekti/määruse ja öeldisverbi vahel võib kasutada rohkem adverbe.

Nimetatud kaks seaduspärasust väljenduvad kolmes morfosüntaktilises mustris:

a) substantiiv ainsuse nominatiivis (subjekt) + adverb + finiitverb indikatiivis ainsuse 3. pöördes (*talumees muudkui muheles*), harva *ma*-infinitiiv (*<tundus> aeg möödas olema*) ja eitava kõne preesens (*jutt enam ei <sobinud>*);

b) substantiiv ainsuse genitiivis (totaalobjekt) + adverb + mineviku liitaja koosseisusu kuuluv *nud*-partitsiip (abiverb jääb mustri vasakkonteksti),

nt *mehe üle võtnud*, *inimese ära tapnud*. Harva on substantiiv ka partsiaalobjekt ainsuse partitiivis – sel juhul kasutatakse predikatiivset *da*-infinitiivi (*meest ärkvele raputada*);

c) substantiiv ainsuse komitatiivis käändsõnalise viisi- või kaasnevusmäärusena + verbipartikkel + põhiverb *nud*-partitsiibi vormis (*<oli> kiiruga alla jooksnud*) – mineviku liitaja vormi kuuluv abiverb jääb SDV-struktuuri vasakkonteksti.

5.3.2. Näiteid adverbita mustritest

Levinud adverbita mustrid on JSV ja JPV (alamklass J-V) ning SVV, SKV ja SSV (alamklass S-V). Kuigi nende esinemus pole võrreldav adverbi sisaldavatega, ei tähenda see, nagu oleks nende roll tekstiloomes väheoluline.

1) JSV-l (nagu ka adverbi sisaldaval JDV-l) on tekstis sidususfunktsioon, kuid erinevalt JDV-st, mida kasutatakse ainult rinnastusega, võib JSV-d kasutada nii rinnastuse (*ja*, *aga*, *ning*, *kuid*) kui ka alistusega (*et*, *kuni*, *kuigi*, *sest*, *kui*, *nagu*). Domineerivad sidesõnad on *ja* (*ja kukk laulis*, *Ja kämblad on*) ning *et* (*et sekretär on* (*oli*), *et riided ei <ole, olnud>*).

JSV morfosüntaktiline varieerumine on eriti keeruline: substantiiv ainsuse, harvem mitmuse nominatiivis on subjekti funktsioonis (*ja hobune tiirles*, *sest Robi ütles*), inessiiv näitab kohta (*ja kõrvades ei <vilista>*), adessiiv ja translatiiv märgivad aega (*Ja hommikulgi oli*, *ja hetkeks vilksatas*), partitiiv tähistab objekti (*ning häält tasandamata*) ja partsiaalsubjekti (*et tegemist on*). Verbi vormikasutus piirdub indikatiivi imperfekti ainsuse (mitmuse) 3. pöördega (*ja seltskond läks*), preesens esineb harvem (*et inimene on*, *Aga kingad on*, *et vaenlased ei <olnud>*).

2) JPV peamine funktsioon on siduda osalauseid (*Sekretär lausus, et see on võimatu*), lause alguses aga avada või täpsustada seda, mis eespool kõneks (*Ja mina ei teinud katsetki teda ümber veenda*) – analoogia JDV ja JSV-ga. Valdavalt kasutatakse alistusseost (*et*, harvem *kui(gi)*, *sest* ja *nagu*), rinnastust esineb umbes neljandiku võrra vähem (*ja*, harvem *kuid*, *aga*, *vaid*, *ent*, *või*).

JSV ja JPV keskmise komponendi vormilise erinevuse taga on erinev funktsionaalne potentsiaal ja morfosüntaktilise varieerumise piirid.

a) Alistusseosega kinnistunud mall sisaldab ainsuse 3. isiku personaalpronoomenit ja verbi indikatiivi preesensi ainsuse 3. pöördes (*et ta on*, *et ta näeb*). Selles mustris varieerub sidesõna (*et*, *kui*, *sest*). Pronoomeni varieerumine toob esile teise kinnistunud malli, milles demonstratiivpronoomen *see* esineb koos verbi imperfektivormi, harvem preesensiga – *et see oli* ja *et*

(sest) see on. Harva kasutatakse personaalpronoomeneid *nad* (*et nad on*, *sest nad lendasid*) ja *ma* (*et ma ei*) subjekti funktsioonis ning *ta* ja *ma* adessiivis adverbiaalina, nt *et tal jääb*, *et tal ei* ja *et mul ei*, *et mul pole*. Varieeruda võib ka kõneviis, indikatiivi kõrval kasutatakse üsna sageli konditsionaali preesensit, nt *et ta oleks (poleks)*, *et see juhtuks*, *et te oleksite*, *et mina oleksin*. Sel juhul on subjekti funktsioonis kasutatud pronoomenite varieerumine üsna avar.

b) Rinnastusseosega on pronoomeni varieerumine päris lai: lisaks ainsuse 3. isikule (*ja ta märkas*, *vaid ta tahtis*) kasutatakse ainsuse 1. ja mitmuse 3. isikut (*ja ma ütlesin*, *aga ma olin*; *ning me sammusime*), harva ka demonstratiivpronoomenid *see* ja *too* (*ent see oli*, *kuid too seisis*). Teine, tunduvalt harvemini kasutatud koordinatiivse seose mall tuleb esile eitava kõne preesensiga + ainsuse 3. isiku personaalpronoomeniga, nt *kuid (aga) ta ei <ütle>*, *ja ta ei <taha>*. Pronoomeni käände varieerumine on seotud kogejaomaja tähistamisega: *Ja mulle näis*, *Kuid minule oli*; *ja tal tekib*, *aga tal on*.

3) SVV esinemus on väike, kuid mustril on tekstiloomes kindel funktsioon – sõnastada hüpoteetilist laadi üldistavaid mõtteavaldusi. Mustrit kasutatakse nii eitavas kui ka jaatavas kõnes, abiverb on konditsionaali preesensis, põhiverb järgneb mustri paremkontekstis, nt *Naine ei oleks <pidanud seda taluma>*, *<ilma et> lausetki öelda oleksin <jõudnud>*. SVV puhul tuleb silmas pidada, et paremkontekstis asuvat põhiverbi võib abiverbist eraldada üks või mitu laiendit, millel on erinevaid funktsioone.

4) SKV on tagasõnafraasi muster, milles adpositsioonide abil tähistatakse paigalolekut ja kohta (*juures*, *all*, *ees*, *vahel*, *sees*, *kohal*, *vastas*, *küljes*, *peal*). Tegevuste mõtestamine oleneb vaatepunktist: tegevused kulgevad millegi-kellegi suunas (*alla*, *ette*, *taha*, *poole*, *otsa*, *kõrvale*) või lähtuvad millestki-kellestki (*alt*, *juurest*). Näiteks: a) koht – *peegli ees seisatas*, *kuuskede all kasvas*; b) millegi-kellegi suunas – *ava poole lendas*, *kupli otsa ronides*; c) millestki-kellestki lähtudes – *Jooriku juurest tulema*, *ukse alt immitses*. Leidub vaid üksikuid aja- (*nädala jooksul oli*, *<kaks> tundi tagasi oli*) ja põhjusmäärusi (*asja pärast sekeldama*). SKV-s eelistatakse finiitöeldist indikatiivi imperfekti ainsuse 3. pöördes (*maja juures oli, käänaku taha kadus, silmade ette kerkis*) ja predikatiivset *ma*-infinitiivi (*tulekuma poole astuma*, *päikese ette seisma*, *Elleni otsa vaatama*). Niisiis on eestikeelses tekstis verbist vasakul kinnistunud leksikaalselt piiratud kohatähenduslik tagasõnafraas, mille kasutuse määravad paigalolek ja vaatepunkt, millest liikumist kirjeldatakse. Tegemist on hästi selge kasutusreegliga.

5) SSV alustab tavaliselt lauset, esimene järjestikustest substantiividest on eestäiend ja teine subjekt, verbi kasutatakse harilikult indikatiivi imperfekti ainsuse 3. pöördes (*Metsa vari ulatus*, *Mehe vaim oli*), harvem mitmuse 3. pöördes (*Laste voodid olid*). SSV on tüüpiline kirjeldavale esituslaadile.

Kuigi adverbita mustrite osakaal verbist vasakul on tekstiloomes tunduvalt väiksem kui adverbi sisaldavatel mustritel, leidub neil rida konkreetseid tekstilisi funktsioone, mille puhul komponentide leksikaalsemantiline ja morfosüntaktiline varieerumine on selgelt piiratud: sidesõnaga algavad mustrid seovad osalauseid ja lauseid tekstiliseks tervikuks; tagasõnafraasi kasutus rajaneb paigaloleku ja liikumissuuna markeerimisel; lause alguses esinev muster, mis koosneb kahest järjestikusest substantiivist (esimene eestäiendina) ja verbist, on iseloomulik kirjeldavale esituslaadile.

6. Järeldusi

Normikohase kirjakeele alusel verbist vasakul ja paremal leitud mustrid on eesti keelele omased (vt punktid 5.1, 5.2). Nende tekstikasutust iseloomustab kindel morfoloogiline struktuur, neid iseloomustavad kindlad leksikaalse, morfosüntaktilise ja funktsionaalse varieerumise piirid (vt punktid 5.3.1 ja 5.3.2). Osa mustritest sisaldab leksikaalgrammatilist dominanti, mõni muster on tekstikasutuses rohkem kinnistunud; üks muster eristub süntaktiliste funktsioonide varieeruvusega, teise alusel kasutatakse ja moodustatakse analüütilisi verbe, kolmandad moodustavad häälduslikult ja morfosüntaktiliselt tõmbuvaid liitseid üksusi jne. Mustrite komponendid võivad olla avara leksikaalsemantilise varieeruvusega, moodustada vabu sõnaühendeid, mis tulevad mustrina esile vaid tänu sellele, et on kasutatud kinnistunud morfoloogilise struktuuri alusel välja kujunenud morfosüntaktilise varieerumise tingimustes. Kõikide mustrite paiknemine kas verbi vasak- või paremkontekstis pole juhuslik, selle põhjal kujunevad eesti keele diskursuspõhised sõnajärjemallid.

Normeeritud kirjakeele kasutusmustrid, nende ainukordne, kinnistunud või varieeruv sõnavara on allikmaterjal aktiivset tüüpi sõnastike ja õpikute koostajatele. Pedagoogilisel eesmärgil saab järjestada mustrite omandamist nende morfosüntaktilise ja leksikaalsemantilise keerukuse ning tekstifunktsioonide alusel. Klastrileidjat, Sketch Engine'i jt analoogseid programme kasutades võib iga vähegi tehnoloogiat valdav inimene iseseisvalt otsida ja leida nt poolametliku kirjavahetuse tüüpilisi sõnastusi ning neid oma eesmärkidel ära kasutada. Leitud mustrite põhjal konkretiseeruvad eesti keele kasutusstandard ja -reeglid. Selles seisneb nii kasutuspõhise keelekäsitluse rakenduslik (sh pedagoogiline, keeletehnoloogiline) perspektiiv kui ka avastuslik väärtus. Uurija

intuitsioonist ja teoreetilistest seisukohtadest sõltumata leitud lingvistiliste mustrite kvalitatiivne analüüs näitab lingvistilisi nähtusi uudses aspektis, avades huvipakkuvaid seoseid (kategoriaalseid, funktsionaalseid, leksikaal-semantilisi, morfosüntaktilisi jm) lekseemide ning vormide kombineerimisel terviktekstis. Võrdlev sünkroonne ja diakroonne keelekasutusmustrite analüüs toob esile kvantitatiivseid ja kvalitatiivseid nihkeid, lingvistilisi arenguid jpm, kinnitades kokkuvõttes põhimõttelist arusaama, et mistahes süsteem on vaid sedavõrd süsteem, kuivõrd selgelt on väljendunud selle asüsteemne olemus kahe vastandliku protsessi – kinnistumise ning vaba varieerumise – tasakaalustatud toimimises. Suletud süsteemid pole arenguvõimelised ja kaovad, mis on oluline lingvistiliste protsesside mõistmiseks.

Lingvistilise klasteranalüüsi hierarhiast tulenev süstemaatika on põhimõtteliselt uut laadi kasutusgrammatilise eesti keele kirjelduse alus. Tundugu see pealegi lihtne, kuid klasteranalüüsi objektipõhine otsing ja rühmitamine formaalsete tunnuste (morfo- ja süntaksimärgendid) distributsiooni alusel mõjutab ühel või teisel moel seda, kuidas me keelest mõtleme, keele olemust seletame (metodoloogia) ja kuidas võiksime seletada. Tegu pole mitte ainult meetodiga, vaid empiirikast lähtuva, induktsioonil rajaneva avastusliku suunaga, mille uurimistulemustel on lisaks rakenduslikele ka iseseisev teoreetiline väljund.

Kommentaarid

[1] Vt Marju Ilves. *Algaja keelekasutaja*. A2-taseme eesti keele oskus. Tallinn: EKS, 2008; Anu-Reet Hausenberg, Marju Ilves, Annekatrin Kaivapalu, Krista Kerge, Katrin Kern, Mare Kitsnik, Ingrid Krall, Karin Rummo, Tiina Rüütmaa. *Iseseisev keelekasutaja*. B1- ja B2-taseme eesti keele oskus. Tallinn: Riiklik Eksami- ja Kvalifikatsioonikeskus, 2008; Krista Kerge. Vilunud keelekasutaja. C1-taseme eesti keele oskus. Tallinn: EKS, 2008.

[2] Vt Anu-Reet Hausenberg, Tiina Kikerpill, Maia Rõigas, Ülle Türk. *Keeleoskuse mõõtmine*. Käsiraamat. Tallinn: TEA Kirjastus, 2003.

[3] Vt ISK = *Iso suomen kielioppi* 2004 (http://scripta.kotus.fi/visk/etusivu.php – 10. jaanuar 2017).

[4] Kahesilbiliste verbide alusel on ajalooliselt kujunenud välja a) tavalised ja b) kontraheerunud kahesilbilised sõnad, millel on vokaaltüvi spirandi kao tõttu ühe silbi võrra lühenenud. <...> *näiteks: *hakkaδan > hakkan, *lükkäδämähan > lükkama, kasteγellen > kastele, *honehessa > hônes, *surnuδeksi > surnuks. Enamik lihttüvesid oli läänemeresoome algkeeles kahesilbilised ja lõppes a-, ä-, e- või e-ga. i-, o-, u-, ü-ga lõppevad kahesilbilised tüved on tuletusliitega sõnatüved või laentüved. Kolmesilbilised sõnatüved olid samuti tuletusliitega sõnatüved või hilised laentüved* (Rätsep 1982: 4–5).

[5] ICLE – The International Corpus of Learner English (https://www.uclouvain.be/en-cecl-icle.html – 20. august 2017).

[6] Vt http://www.keeleveeb.ee (9. juuni 2017).

[7] Vt http://www.cl.ut.ee/ressursid/sagedused/index.php?lang=et (9. juuni 2017).

[8] Vt http://www.folklore.ee/justkui/sonastik/ (9. juuni 2017).

[9] Vt http://www.cl.ut.ee/ressursid/pysiyhendid/index.php?lang=et (9. juuni 2017).

[10] Vt http://www.eki.ee/dict/sp/ (9. juuni 2017) ja http://mobile.dspace.ut.ee/bitstream/handle/10062/50084/sonamoodustus_kasik.pdf (9. juuni 2017). Silvi Vare (2012) eesti sõnaperede sõnastik toob esile derivatiivsed pesad, avab astmeliselt iga pesa sees algtüve derivatiivse potentsiaali, pakub liitsõnade moodustusmallid, sõna grammatilise struktuuri, tüüpväljendid jm. Kasutajaliidese abil on võimalik leida sõnalõpuanaloogiat ja morfoloogilisi seoseid, millel on otsene rakendus õppematerjalide koostamisel ja praktilises keeleõppes. Reet Kasik (2015) kirjeldab sõnamoodustust ühtsel funktsionaalsemantilisel alusel sõnaliikide kaupa, igas sõnaliigis tähenduste järgi koos tüüpiliste ja harvem kasutatavate tuletusliidete ning näidetega. Sõnamoodustusraamatust saab kätte eestikeelses tekstikasutuses samalaadse malli alusel moodustatud sõnu, mida omandatakse analoogia põhjal.

[11] Vt Della Summers (toim). *Longman Dictionary of Contemporary English*. 4. tr. Longman, 2003. Sisaldab 106 000 sõna ja fraasi, 220 000 sõnakombinatsiooni; lisatud CD-ROM.

[12] Igor Mel'čuk, Aleksandr Zholkovsky. Explanatory combinatorial dictionary of modern Russian. *Semantico-syntactic Studies of Russian Vocabulary*. Viin: Wiener Slawistischer Almanach, 1984. (Vt lisaks Mel'čuk 1995: 5.)

[13] Vt http://www.eki.ee/dict/psv/ (18. mai 2017).

[14] Vt Marju Ilves. *Algaja keelekasutaja*. A2-taseme eesti keele oskus. Tallinn: EKS, 2008. Lisa 2. Esmane sõnastik, lk 137 jj.

[15] Vt Heiki-Jaan Kaalep, Kadri Muischnek. *Eesti kirjakeele sagedussõnastik*. Tartu: TÜ Kirjastus, 2002.

[16] Vt Mall Laur. *Esmane verstapost*. Eesti keele suhtluse algtase. Tallinn: REKK, 1998.

[17] Vt Hele Pärn, Leeni Simm. *Eesti keele baassõnastik*, esmatrükk 1988; sisaldab venekeelseid tõlkevasteid. Jätkuväljaannetes on lisatud soome-, rootsi- ja ingliskeelsed vasted.

[18] Vt Evi Šteinfeldt. *Russian Word Count: 2500 Words Most Commonly Used in Modern Literary Russian*. Guide for Teachers of Russian. Moskva: Progress Publishers, 1965. Sõnastik pole oma aktuaalsust tänaseni kaotanud (vt Laleko 2010). Koolisõnastik on valminud koostöös Enda Roovetiga: Энда Роовет, Эви Штейнфельдт. *Словарь-минимум русского языка для 2-8 классов эстонских школ*. Таллинн: Валгус, 1965.

[19] http://evkk.tlu.ee/Marks/global_marks/marks_public.html (19. august 2017).

[20] Arvutatud valemi $k \approx \sqrt{n \div 2}$ alusel, kus k on optimaalsuskoefitsient ja n tähistab trigrammide hulka. Klastrileidja eraldas ilukirjanduskeele valimis kokku 52 071 trigrammi, neist 42 183 on ainukordsed ja verbilõpulisi on 11552 (k = 11552 : 2 = 5776, millest ruutjuur on 76).

[21] Verbiga algavaid trigramme on 8184 (k = 8184 : 2 = 4092, millest ruutjuur on 63,9687.~ 64).

Kirjandus

Aedmaa, Eleri 2016. Eesti keele ühendverbide kompositsionaalsuse määramine. *Eesti Rakenduslingvistika Ühingu aastaraamat* 12, lk 5–23 (doi:10.5128/ERYa12.01).

Aedmaa, Eleri 2015. Statistilised meetodid ühendverbide tuvastamisel tekstikorpusest. *Eesti Rakenduslingvistika Ühingu aastaraamat* 11, lk 37–54 (doi:10.5128/ERYa11.03).

Allkivi, Kais 2016a. *C1-tasemega eesti keele õppijate kirjalik keelekasutus võrdluses emakeelekõnelejatega: samalaadsusi ja nihkeid verbist paremal paiknevas kontekstis*. Magistritöö. Tallinn: Tallinna Ülikool (http://www.etera.ee/zoom/20076/view?page=1&p=separate&view=0,299,2481,1414 – 2. oktoober 2017).

Allkivi, Kais 2016b. C1-tasemega eesti keele õppijate ja emakeelekõnelejate kirjaliku keelekasutuse võrdlus verbialguliste tetragrammide näitel. *Lähivõrdlusi* 26, lk 54–83 (doi: 10.5128/LV26.02).

Allkivi, Kais & Eslon, Pille & Trainis, Jekaterina 2017. Kontseptuaalselt sidusa mõistevara kujunemine: hierarhiline klasteranalüüs. *16. rakenduslingvistika kevadkonverents, 20.–21.04.2017*. Teesid. Eesti Rakenduslingvistika Ühing, 28–29 (https://www.rakenduslingvistika.ee/wp-content/uploads/2016/04/Teesid-2017-3.pdf – 24. august 2017).

Apresjan *et al.* 2010 = Apresian, Valentina & Apresian, Iurii & Babaeva, Elizaveta & Boguslavskaia, Ol'ga & Iomdin, Boris & Krylova, Tat'iana & Levontina, Irina & Sannikov, Andrei & Uryson, Elena 2010. Iurii Derenikovich Apresian (vast toim). *Prospekt aktivnogo slovaria russkogo iazyka*. Moskva: Iazyki slavianskikh kul'tur.

Apresjan *et al.* 2006 = Apresian, Valentina & Apresian, Iurii & Babaeva, Elizaveta & Boguslavskaia, Ol'ga & Iomdin, Boris & Krylova, Tat'iana & Levontina, Irina & Sannikov, Andrei & Uryson, Elena 2006. *Iazykovaia kartina mira i sistemnaia leksikografiia*. Moskva: Iazyki slavianskikh kul'tur.

Bartning, Inge & Martin, Maisa & Vedder, Ineke (toim) 2010. *Communicative proficiency and linguistic development: intersections between SLA and language testing research*. EUROSLA Monographs Series 1 (http://eurosla.org/monographs/EM01/EM01tot.pdf – 5. november 2017).

Biber, Douglas & Conrad, Susan & Reppen, Randi 2006. *Corpus Linguistics. Investigating language structure and use*. New York: Cambridge University Press.

Chomsky, Noam 1957. *Syntactic Structures*. Haag/Pariis: Mouton.

Conklin, Kathy & Schmitt, Norbert 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29 (1), lk 72–89 (doi: 10.1093/applin/amm022).

Eslon, Pille 2017. Keelekasutusmustrid verbist paremal: morfosüntaktiline ja leksikaalsemantiline varieerumine. *Lähivõrdlusi* 27, lk 17–64 (doi:10.5128/LV27.00).

Eslon, Pille 2014a. Adverbi sisaldavate struktuuride tekstifunktsioonidest eesti ilukirjandus- ja õppijakeeles. *Lähivõrdlusi* 24, lk 15–46 (doi:10.5128/LV24.01).

Eslon, Pille 2014b. Morfosüntaktilise ja leksikaalse varieerumise piiridest: ilukirjandus- ja õppijakeele kasutusmustrite võrdlus. *Eesti Rakenduslingvistika Ühingu aastaraamat* 10, lk 55–71 (doi:10.51.28/ERYa10.04).

Eslon, Pille 2013. Kahe keelekasutusvariandi võrdlus: morfoloogilised klassid ja klastrid. *Lähivõrdlusi* 23, lk 13–38 (doi:10.5128/LV23.01).

Eslon, Pille & Paeoja, Heleriin 2015. Samatähenduslike sünteetiliste ja analüütiliste verbide kasutamine. *Lähivõrdlusi* 25, lk 63–104 (doi:10.5128/LV25.04).

Greenberg, Joseph 1990. A relation of frequency to semantic feature in a case language (Russian). Denning, Keith & Kemmer, Suzanne (toim). *On Language. Selected Writings of Joseph H. Greenberg*. Stanford, CA: Stanford University Press, lk 207–226.

Hint, Mati 2016. Mõõtmised ei loo teooriat. *Keel ja Kirjandus* 8–9, lk 627–637.

Holl, Alfred & Behrschmidt, André & Kühn, Alexander 2004. *Rückläufige Register zur russischen und deutschen Verbalmorphologie: Aufbereitung mit Datenanalyseverfahren der Informatik (Data Mining)*. Studia et exempla linguistica et philologica. Regensburg: Roderer.

Jarvis, Scott & Crossley, Scott A. (toim) 2012. *Approaching language transfer trough text classification: Exploring in the detection-based approach*. Bristol/Buffalo/Toronto: Multilingual Matters.

Jürine, Anni 2016. *The Development of Complex Postpositions in Estonian: A Case of Grammaticalization and Lexicalization*. Dissertationes Philologiae Estonicae Universitatis Tartuensis 38. Tartu: Tartu Ülikooli Kirjastus.

Kaalep, Heiki-Jaan 2015. Eesti verbi vormistik. *Keel ja Kirjandus* 1, lk 1–15.

Kaalep, Heiki-Jaan 2012. Käänamissüsteemi seaduspärasused. *Keel ja Kirjandus* 6, lk 418–449.

Kaalep, Heiki-Jaan & Muischnek, Kadri 2009. Eesti keele püsiühendid arvutilingvistikas. *Eesti Rakenduslingvistika Ühingu aastaraamat* 5, lk 157–172 (doi: 10.5128/ERYa5.10).

Kaivapalu, Annekatrin 2013. Kielten järjestelmien vertailusta kieltenvälisen vaikutuksen tutkimukseen. Kolehmainen, Leena & Miestamo, Matti & Nordlund, Taru (toim). *Kielten vertailun metodiikka*. Helsinki: Suomalaisen Kirjallisuuden Seura, lk 293–323.

Kaivapalu, Annekatrin & Martin, Maisa 2014. Measuring perceptions of cross-linguistic similarity between closely related languages: Finnish and Estonian noun morphology as a testing ground. Paulasto, Heli & Meriläinen, Lea & Riionheimo, Helka & Kok, Maria (toim). *Language Contacts at the Crossroads of Disciplines*. NewCastle Upon Tyne: Cambridge Scholars Publishing, lk 283–318.

Kallas, Jelena 2013. *Eesti keele sisusõnade süntagmaatilised suhted korpus- ja õppeleksikograafias*. Tallinna Ülikooli humanitaarteaduste dissertatsioonid 32 (http://e-ait.tlulib.ee/303/1/kallas_jelena.pdf – 12. mai 2017).

Kallas, J *et al.* = Kallas, Elena & Koppel', Kristina & Kallas, Roman 2017. Avtomaticheskoe sostavlenie slovaria kollokatsii na osnove korpusa. Trudy mezhdunarodnoi konferentsii "Korpusnaia lingvistika-2017", 27–30 iiunia 2017. Sankt-Peterburg: Sankt-Peterburgskii gosudarstvennyi universitet, lk 195–200.

Kallas, Jelena & Koppel, Kristina & Tuulik, Maria 2015. Korpusleksikograafia uued võimalused eesti keele kollokatsioonisõnastiku näitel. *Eesti Rakenduslingvistika Ühingu aastaraamat* 11, lk 75–94 (doi: 10.5128/ERYa11.05).

Kallas, Jelena & Tiits, Mai & Tuulik, Maria & Koppel, Kristina & Jürviste, Madis 2014. *Eesti keele põhisõnavara sõnastik*. Tallinn: EKS. (http://www.eki.ee/dict/psv/ – 18. august 2017).

Kallas, Jelena & Tuulik, Maria & Jürviste, Madis 2012. Leksikograafilise tarkvara Sketch Engine eesti keele moodul. *ESUKA* [Eesti ja Soome-ugri Keeleteaduse Ajakiri] 3–2, lk 57–77.

Kasik, Reet 2015. *Sõnamoodustus*. Eesti keele varamu I. Tartu.

Klavan, Jane 2012. *Evidence in Linguistics: Corpus-Linguistic and Experimental Methods for Studying Grammatical Synonymy*. Dissertationes Linguisticae Universitatis Tartuensis 15. Tartu: University of Tartu.

Koppel, Kristina & Kallas, Jelena 2016. Õppijasõbralik korpuslause: automaatse valiku võimalusi. *Lähivõrdlusi* 26, lk 222–250 (doi: 10.5128/LV26.07).

Küngas, Annika 2014. *Pragmaatiliste markerite kujunemine ja funktsioonid eesti keeles lt-sõnade näitel*. Dissertationes Philologiae Estonicae Universitatis Tartuensis 36. Tartu: Tartu Ülikooli Kirjastus (http://dspace.ut.ee/bitstream/handle/10062/42733/kungas_annika.pdf?sequence=1 – 12. mai 2017).

Küngas, Annika 2013. *Põhimõtteliselt* või *praktiliselt Paides* – kahe sarnase funktsiooniga sõna käeäigust. *Eesti Rakenduslingvistika Ühingu aastaraamat* 10, lk 209–226 (doi:10.5128/ERYa10.13).

Laleko, Oksana 2010. On Covert Tense-Aspect Restructuring in Heritage Russian: A Case of Aspectually Transient Predicates. Iverson, Michael *et al.* (toim). *Proceedings of the 2009 Mind / Context Divide Workshop*. Somerville, MA: Cascadilla Proceedings Project, lk 72–83.

Levenshtein, Vladimir 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady* 10 (8), lk 707–710 (https://nymity.ch/sybilhunting/pdf/Levenshtein1966a.pdf – 8. aprill 2017).

Liin, Krista 2009. Komavigade tuvastaja. Eslon, Pille & Õim, Katre (toim). *Korpusuuringute metodoloogia ja märgendamise probleemid*. Tallinn: Tallinna Ülikooli Kirjastus, lk 99–114.

Lyytikäinen, Erkki 2012. Erään frekventatiivijohdostyypin produktiivisuudesta. *Virittäjä* 1, lk 114–117.

Matsak, Erika & Eslon, Pille & Kippar, Jaagup 2010. Eesti keele sõnajärje vealeidja prototüübi arendamine. Eslon, Pille & Õim, Katre (toim). *Korpusuuringute metodoloogia ja märgendamise probleemid*. Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 12, lk 59–100.

Mel'čuk, Igor 1995. Semantics of two emotion verbs in Russian: *bojat'sja* '[to] be afraid' and *nadejat'sja* '[to] hope'. Mel'chuk [Mel'čuk], Igor'. *Russkii iazyk v modeli 'smysl <=> tekst'*. Moskva – Vena: Iazyki russkoi kul'tury, lk 81–133.

Metslang, Helena & Matsak, Erika 2010. Kesksete lausekomponentide järjestus õppijakeeles: arvutianalüüsi katse. *Eesti Rakenduslingvistiks Ühingu aastaraamat* 6, lk 175–194 (doi:10.5128/ERYa6.11).

Määttä, Tuija 2012. Oppikirjan sanaston vaikutuksesta ruotsinkielisten alkeistason suomenoppijoiden kirjallisiin tuotoksiin. *Lähivõrdlusi* 22, lk 188–218.

Ots, Sander 2012. *Statistikapõhise tarkvara loomine morfoloogiliste kollokatsioonide eraldamiseks eesti keele tekstidest*. Bakalaureusetöö. Tallinna Ülikooli informaatika instituut.

Ots, Sander 2011. *Tarkvara statistiliste kollokatsioonide eraldamiseks ning selle rakendus morfosüntaktilises analüüsis*. Seminaritöö. Tallinna Ülikooli informaatika instituut.

Paeoja, Heleriin 2015. *Analüütiliste / sünteetiliste verbipaaride kasutusmustrid 1990ndate aastate eesti ilukirjanduskeeles*. Magistritöö. Tallinna Ülikooli eesti keele ja kultuuri instituut.

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

Remes, Hannu 2009. *Muodot kontrastissa. Suomen ja viron vertailevaa taivutusmorfologiaa*. Acta Universitatis Ouluensis B 90. Oulu: Oulu Ülikool.

Ruutma, Mirjam & Kyröläinen, Aki-Juhani & Pilvik, Maarja-Liisa & Uiboaed, Kristel 2016. Ambipositsioonide morfosüntaktilise varieerumise kirjeldusi kvantitatiivsete profiilide abil. *Keel ja Kirjandus* 2, lk 92–113.

Rätsep, Huno 1982. *Eesti keele ajalooline morfoloogia I*. 2., parand ja täiend tr. Tartu: Tartu Riiklik Ülikool.

Saukkonen, Pauli & Haipus, Marjatta & Niemikorpi, Antero & Sulkala, Helena 1979. *Suomen kielen taajuussanasto*. Porvoo/Helsinki/Juva: Söderström.

Sirts, Kairit 2012. Noisy-Channel Spelling Correction Models for Estonian Learner Language Corpus Lemmatisation. *Human Language Technologies – The Baltic Perspective*. IOS Press, lk 213–219 (doi:10.3233/978-1-61499-133-5-213).

Sirts, Kairit & Võhandu, Leo 2009. Korpuste tükeldamine: rakendusi silpide ning allkeeltega. *Eesti Rakenduslingvistika Ühingu aastaraamat* 5, lk 251–266.

Šajkevitš 1976 = Shaikevich [Šajkevitš], Anatolii 1976. Distributivno-statisticheskii analiz v semantike. *Printsipy i metody semanticheskikh issledovanii*. Moskva: Nauka, lk 353–378.

Ševtšenko, Marina 2014. *Eesti keele kui teise keele 8. klassi õpiku temaatiline sõnavara ja grammatika*. Magistritöö. Tallinna Ülikooli eesti keele ja kultuuri instituut.

Tognini-Bonelli, Elena 2001. *Corpus linguistics at work*. Studies in korpus linguistics 6. Amsterdam/Philadelphia: John Benjamins Publ. Co.

Trainis, Jekaterina 2015. Linguistic cluster analysis: A method for describing language units and indicating regularities in language. Malec, Wojciech & Rusinek, Marietta (toim). *Within language, beyond theories*. Vol. III. Discourse analysis, pragmatics and corpus-based studies. Newcastle Upon Tyne: Cambridge Scholars Publishing, lk 229–243.

Trainis, Jekaterina & Allkivi, Kais 2014. Ilukirjanduskeelest uue pilguga. *Eesti Rakenduslingvistika Ühingu aastaraamat* 10, lk 283–306 (doi: 10.5128/ERYa10.18).

Tšernõšuk, Anna 2016. *Eesti keele kui teise keele 9. klassi õpiku temaatiline sõnavara*. Bakalaureusetöö. Tallinna Ülikool.

Uiboaed, Kristel 2013. Kollostruktsioonilised meetodid ja konstruktsioonilise varieerumise tuvastamine. *ESUKA* [Eesti ja Soome-ugri Keeleteaduse Ajakiri] 4–1, lk 185–204 (doi: 10.12697/jeful.2013.4.1.11).

Vainik, Ene 2017. Eesti keele assotsiatsioonisõnastik. *16. rakenduslingvistika kevadkonverents, 20.–21.04.2017*. Teesid. Eesti Rakenduslingvistika Ühing, 17 (https://www.rakenduslingvistika.ee/wp-content/uploads/2016/04/Teesid-2017-3.pdf –24. august 2017, tutvustust vt http://www.eki.ee/~ene/kodanikuteadus/assotsiatsioonid.html – 6. september 2017).

Vainik, Ene 2016. *Eesti tunded. Sõnaportreed*. Tallinn: EKSA.

Vajjala, Sowmya & Lõo, Kaidi 2014. Automatic CEFR level prediction for Estonian learner text. *Proceedings of the third workshop on NLP for computer-assisted language learning*. NEALT Proceedings Series 22 / Linköping Electronic Conference Proceedings 107, lk 113–127 (http://www.aclweb.org/anthology/W14-3509 – 15. august 2017).

Vajjala, Sowmya & Lõo, Kaidi 2013. Role of morpho-syntactic features in Estonian proficiency classification. *Proceedings of the 8th Workshop on Innovative Use of NLP for Building Educational Applications* (BEA8). Association for Computational Linguistics.

Valdmets, Annika 2010. *Mõne modaalpartikli kujunemine ja kasutamine eesti kirjakeeles alates 1890. aastast*. Magistritöö. Tartu Ülikool (http://www.murre.ut.ee/arhiiv/naita_pilt.php?materjal=kasikiri&materjal_id=D1625&sari=D –18. mai 2017).

Vare, Silvi 2012. *Eesti keele sõnapered. Tänapäeva eesti keele sõnavara struktuurianalüüs*. 1. ja 2. kd. Tallinn: EKS.

Widdowson, Henry G. 2000. The limitations of linguistics applied. *Applied Linguistics* 21 (1), lk 3–25.

Õim, Katre 2014. *Metafoorsete sõnaühendite automaatse tuvastamise probleeme*. Ettekanne Eesti Kognitiivse Keeleteaduse Ühingu 3. aastakonverentsil Tartus 4. aprillil 2014. Ettekande slaidid.

Õim, Katre 2009. Alternatiivseid mooduseid fraseoloogia esitamiseks sõnastikus. Eslon, Pille & Õim, Katre (toim). *Korpusuuringute metodoloogia ja märgendamise probleemid*. Tallinn: Tallinna Ülikooli Kirjastus, lk 136–164.

Summary

Usage-based language description: Linguistic cluster analysis and it's perspectives for pedagogical purposes

Pille Eslon
Tallinn University, School of Digital Technologies, associate professor
pille.eslon@tlu.ee

Keywords: corpus linguistics, Estonian studies, linguistic cluster analysis, usage-based language study

Linking Estonian linguistic proficiency to reference levels of the CEFR and different educational stages does not rely on research but is based on deep-rooted perceptions. More veracious data can be obtained by comparing a native speaker's language usage patterns to morphological and lexical preferences characteristic to speakers of every language level. For this purpose, tools for automatic text processing (which are mainly created on the basis of English) and different techniques for data analysis are needed. The article introduces an original computer program called Cluster Catcher that has been developed in the Tallinn University for finding usage patterns from Estonian written language texts.