doi:10.5128/LV28.01

Teksti keelekasutusmustrid ja lingvistiline klasteranalüüs

PILLE ESLON, KAIS ALLKIVI-METSOJA
Tallinna Ülikool

Ülevaade. Suurte korpuste automaatsel töötlemisel kasutatakse erinevat keeletarkvara ja statistilist analüüsi, mille valik ning kombineerimisvõimalused sõltuvad keelest, uurimisobjektist ja eesmärkidest. Artiklis tutvustame teksti keelekasutusmustrite otsimiseks mõeldud integreeritud tarkvararakendust Klastrileidja ja selle toimesüsteemi, anname ülevaate lingvistilise klasteranalüüsi abil saadud uurimistulemustest. Eesmärk on seletada, mida selle meetodi rakendamine loomuliku keele töötluse käigus võimaldab avastada eesti keele ja õppija keelekasutuse kohta ning kuidas neid teadmisi pedagoogilistel vajadustel rakendada.

Võtmesõnad: loomuliku keele töötlus; keelekasutusmustrid; õppijakeel; eesti keel

1. Klastrileidja – integreeritud tarkvararakendus

Klastrileidja (Ots 2011; 2012) on arendatud Erika Matsaku sõnajärjeleidja prototüübi alusel (Matsak jt 2010; Metslang & Matsak 2010). Programm otsib morfosüntaktiliselt märgendatud eestikeelsetest tekstidest sarnase morfo- ja süntaksimärgendite[1] järgnevusega n-gramme, koondab need ühte rühma ehk klastrisse ning registreerib sageduse. Tegemist on arvutitehnoloogia ja matemaatilise statistika integreeritud

[1] Märgendeid ja nende seletusi vt http://kodu.ut.ee/~kaili/Korpus/pindmine/ (16.2.2017), samuti Muischnek jt 2012: 78.

tarkvararakendusega, mis toob formaalsete tunnuste alusel esile seaduspärasusi keele elementide loomulikus järgnevuses ja kombineerimisvõimalustes.

Kasutajal on vabadus seadistada rakendus oma uurimistöö eesmärkidele vastavalt. Varieerida saab järjendi pikkust: üks komponent – unigramm (mugav võrrelda sõnavormide esinemust nt erinevatel keeleoskustasemetel või allkeeltes, leida aktiivne sõnavara), kaks komponenti – bigramm (nt vabad ja seotud sõnaühendid), kolm komponenti – trigramm (nt analüütiliste verbide kasutusmallid, analüütilised konstruktsioonid), neli komponenti – tetragramm (nt diskursuspõhised aktiivsed sõnajärjemallid). Valida saab ka lingvistiliste objektide vahel – morfoloogia, süntaks, morfosüntaks – ja otsustada, kas mustrite otsingul on vaja arvestada (osa)lause piiriga. Need võimalused räägivad analüüsi paindlikkusest. Näiteks seadistasime Klastrileidja otsima eesti vahekeele korpuse (EVKK) vene emakeelega gümnaasiumiõpilaste eesti keele olümpiaaditööde alamkorpuse esseedest (58 614 sõnet) kolmest komponendist koosnevaid sõnavormide järgnevusi ehk trigramme, varieerides samas lingvistilist uurimisobjekti.

1) Objekt – süntaktilised funktsioonid; tulemus – nt kolm sagedamat süntaktilist struktuuri **CLB @J @SUBJ @+FMV ehk (osa)lausealguline sidend + subjekt + finiitöeldis (451 näidet, nt *et autor tahab*), @SUBJ @+FMV @ADVL ehk subjekt + finiitöeldis + adverbiaal (446, nt *autor kirjeldab mitte*) ja @+FMV @ADVL @ADVL ehk finiitöeldis + adverbiaal + adverbiaal (415, nt *on tänapäeval nii*).
2) Objekt – morfoloogilised kategooriad ehk sõnaliigid; tulemus – kolm sagedamat morfoloogilist struktuuri ehk sõnaliigijärjendit: a) verb + adverb + adverb (64 näidet, nt *on veel vara*); b) eitava kõne marker + verb + adverb (57, nt *ei tule enam*); c) adverb + adverb + adverb (52, nt *juba kusagilt mujalt*).
3) Objekt – grammatilised vormid; tulemus – nt sageduselt teise morfoloogilise struktuuri vormikasutusele on iseloomulik üldeitus koos intransitiivse verbi indikatiivi preesensi finiitvormile

järgneva adverbiaalse laiendiga: _V_ aux neg + _V_ main indic pres ps neg #FinV #Intr + _D_ (nt *ei tule vist*).

4) Objekt – morfosüntaks; tulemus – nt sagedamat morfosüntaktilist struktuuri kirjeldab järgnevus _V_ aux neg @NEG + _V_ main indic pres ps neg #FinV #Intr @+FMV + _D_ @ADVL (nt *ei tule enam*).

Lingvistilise objekti varieerimine annab ettekujutuse erinevatesse keeletasanditesse kuuluvate elementide sagedatest kombineerimisvõimalustest ehk kasutusreeglitest, mida vene lähtekeelega eesti keele olümpiaaditööde kirjutajad on eelistanud: eitav kõneliik, intransitiivse verbi indikatiivi preesensi vorm lihtöeldise funktsioonis. Adverbiaalseid funktsioone süntaksianalüsaator ei täpsusta. (Vt ka Eslon 2017a: 226–228)

Kirjeldatud keelekasutusmustrites peegeldub analüüsiks valitud objekti olemus, sest andmekaeve põhimõttel töötav programm on need objektipõhiselt leidnud ja rühmitanud. Otsing tugineb lingvistiliste andmete formaalsele esitusele ja rakendub automaatselt uurija tahtest või teoreetilistest seisukohtadest sõltumata. Klastrileidja alt-üles toimesüsteemi kohaselt eristuvad sarnase vormistusega morfosüntaktilised järjendid (n-grammid), mis liigituvad suurematesse rühmadesse (klastritesse). Tegemist pole siiski klassikalise, vaid **lingvistilise klasteranalüüsiga**, mida statistilise analüüsiga integreeritud tarkvararakendus võimaldab.

Kuna n-grammide ja klastrite hierarhiline liigendus tuleneb Klastrileidja toimesüsteemist, siis on see objektiivne alus ka keelekasutusmustrite klassifitseerimiseks ning kvalitatiivse lingvistiline analüüsi tulemuste süsteemseks esituseks. Oma uurimustes (vt Eslon 2014a; 2014b; Trainis & Allkivi 2014; Eslon & Paeoja 2015; Allkivi 2016a; 2016b; Eslon 2017a; 2017b), samuti teistes lingvistilise klasteranalüüsi abil tehtud uurimustes (nt Ševtšenko 2014; Trainis 2015; Paeoja 2015; Tšernõšuk 2016; Trainis 2017; Voolaid 2018) oleme n-grammide ja klastrite hierarhiat laiendanud klassi ning alamklassi mõistetega, sest suure hulga andmete korral tulevad keelekasutustendentsid selgemalt esile kõrgematel üldistuse tasanditel. Niisiis moodustub n-grammide, klastrite, alamklasside ja klasside

põhjal **lingvistilise klasteranalüüsi hierarhia**, mis lubab Klastrileidja alt-üles toimimisel saadud tulemusi klassifitseerida ja süstematiseerida ka vastupidiselt, st ülalt-alla põhimõttel – klassid, alamklassid, klastrid ja n-grammid. Mõlemal juhul on tegemist olemusliku klassifikatsiooniga, mille element on teksti **keelekasutusmuster**, mis leitud objektipõhiselt formaalsete tunnuste põhjal. Selle üksuse alusel üldistuvad lingvistiliste elementide loomulikud kasutusreeglid tekstis, mis omab tähendust nii lingvistika kui ka keeleõppe seisukohalt. Samas tuginetakse pedagoogilistes rakendustes (õpikud) ja õppijakeele uurimustes (nt Kitsnik 2018) abstraktsele **konstruktsiooni** mõistele, mille sisu pole üheselt määratletud. Kerkib küsimus, kas erinevat liiki üksused, mida konstruktsioonidena käsitletakse, on keeleõppe elemendina sobivamad kui teksti keelekasutusmustrid, mis saadud integreeritud tarkvararakenduse abil objektipõhiselt. Kas õpikutes esinevate konstruktsioonide omandamine kujundab suurema võimekuse kirjalikke tekste produtseerida ja suulises suhtluses tulemuslikumalt osaleda kui lähtumine keelekasutusmustritest?

2. Teksti keelekasutusmustrid *vs.* konstruktsioonid

Teksti keelekasutusmustrid ja konstruktsioonid on erineva lingvistilise tasandi üksused. Konstruktsioon on **lause element**, konstruktsiooni element on sõnavorm, mille grammatilised funktsioonid sõltuvad teda ümbritsevatest sõnavormidest (nt *sinine taevas* – omadussõna eestäiendi funktsioonis ühildub nimisõnaga arvus ja käändes, tähistades põhisõna tunnust) ja nende vabast varieerumisest (*pilvine*, *vihmane*, *päikeseline* jne *taevas*) või on piiratud semantiliselt sobivate kooslustega (*kisendav ülekohus*, *kange kohv*) ja idiomaatiliste üksustega (*pani pihta*, *viskas varvast*).

Keelekasutusmuster on **teksti element**, millel kindlad tekstifunktsioonid. Näiteks: ajatähenduslik stereotüüp *toimus sel aastal*, *oli sel suvel* seob kirjeldava teksti osad loogiliseks tervikuks, tagades sujuvad üleminekud; perioodilist korduvust tähistav muster *on kord juba*, *tuleb*

kord jälle tähistab kõnesoleva omadust; rinnastav-alistavad rühmsidendid nagu *mitte ainult* (*enam üldse*, *nii palju*) *... , vaid* (*kui*) on kasutusel mööndu se väljendamiseks; järjend tegusõna-määrsõna-partikkel on reeglina ühendverbide kasutusmall verbist paremal (*läks peagi laiali*, *küürutas kähku maha*, *andis niivõrd järele*, *pidas väga kinni*, *tuli jälle tagasi*) jne. (Vt Eslon 2017b)

Konstruktsioonide pikkus on tavaliselt kaks-kolm elementi, sõnavormid kombineeruvad neis kategoriaalsete ja semantilis-süntaktiliste seoste põhjal. Niisugused üksused tulevad esile nt lause puudepangas, võivad olla kirjeldatud fraasi- ning sõltuvusstruktuuridena ja kombineeritud kujul (vt Muischnek jt 2016).

Keelekasutusmustrite pikkus oleneb uurimiobjekti olemusest ja eesmärkidest ning on seega paindlik. Tavaliselt analüüsitakse samuti kahest-kolmest komponendist koosnevaid tekstilisi järgnevusi, millel erinev sõnavara ja vormistik – stereotüüpne, teatud piires kinnistunud või varieeruv. Suure esinemusega mustrid sisaldavad aktiivset sõnavara ja grammatikat, tuues esile keele elementide kommunikatiivse ning funktsionaalse võimekuse.

Konstruktsioonide ja teksti keelekasutusmustrite eristamiseks kasutatakse mitmesuguseid meetodeid, neid kirjeldatakse ja seletatakse erinevatelt teoreetilistelt positsioonidelt ning rakendatakse erinevatel vajadustel. Õppijakeelt võib sihtkeelega võrrelda nii abstraktsete süntaksiskeemide ehk konstruktsioonide kui ka loomulikus keelekasutuses tüüpiliselt kooskasutatud sõnavormide alusel. Mõlemat liiki üksustel on keele süsteemsete seoste ülesehitamise protsessis oma osa ja omad reeglid – konstruktsioonidel sõnaühendi süntaks (keele elementide süteemsed-struktuursed seosed) ja mustritel tekstisüntaks (keele elementide süntagmaatilised seosed).

Näiteks ühe eestikeelsetes tekstides sageli esineva mustri verb-substantiiv-adverb (*pööras pilgu ära*, *küsis Tehvan pahaselt*) kasutusreeglid kujunevad verbi, substantiivi ja adverbi semantilis-süntaktiliste seoste põhjal tekstis, mille spetsiifikat on keeruline esile tuua lause fraasistruktuuride või sõltuvuspuude alusel.

Kõigepealt, öeldisverbi kasutatakse stereotüüpselt indikatiivi imperfekti ainsuse 3. pöördes (*pööras*, *küsis*), mis on mustri morfosüntaktiline dominant. Substantiivi funktsioonid sõltuvad sellest, kas subjektil on elus (*ütles neiu tasa*) või elutu referent (*muutus samm veelgi* <aeglasemaks>), kas substantiivi kasutatakse subjekti (*küsis Tehvan pahaselt*), totaalobjekti (*sai alguse juba* <eile>) või käändelise määrusena (*asus rajoonikeskusest niipalju* <eemal, et ... >). Adverbi funktsioonid ulatuvad verbipartiklist (*pööras pilgu ära*, *võttis buketi vastu*) määruseni (*vastas Lutrin ebamääraselt*).

Teiseks, mida piiratum on kirjeldatud mustris verbi semantika (kasutatakse kõne- ja tuumverbe) ning morfosüntaks (finiitöeldis, imperfekti ainsuse 3. pööre), seda huvitavamaid kasutusreegleid tuleb esile elusat/elutut referenti tähistava substantiivi kooskasutuses adverbiga: a) kõneverbid (*lausus*, *vastas*) + elus referent subjekti funktsioonis (*ajakirjanik*, *Lutrin*) + tavaliselt *lt*-lõpuline adverb (nt *õpetlikult*) viisimäärusena; b) tuumverbid (*läks*, *jäi*) + elutut referenti tähistav substantiiv moodustavad analüütilisi verbe (*läks lukk rikki*, *võttis kaane ära*) koos verbipartikli, kivinenud noomeni vormi või infinitiiviga, mis reeglina on eraldatud põhiverbist ühe või mitme komponendiga (nt subjekti sisaldavates näidetes *oli vastus ka* <käes>, *sirutas mees jälle* <käe välja>, *vaatas kass uuesti* <tagasi>; objekti sisaldavates näidetes *surus käsivarred kõvasti* <kokku>, *pesi kleidi just* 'äsja' <puhtaks>).

Kolmandaks, verbile ja substantiivile järgneva adverbi funktsioonid sõltuvad verbi sünteetilisusest/analüütilisusest. Kui kasutatakse analüütilist verbi, siis on substantiiv totaalobjekti ja adverb reeglina partikli funktsioonis (*võttis buketi vastu*), kui sünteetilist verbi, siis järgneb partsiaalobjekti funktsioonis substantiivile reeglina rõhupartikkel *ka* (*tahtis omakseid ka* <näha>, harvem erinevat liiki määrused ning substantiiv on sel juhul subjekti funktsioonis (*kadus ülempreester kuhugi*, *ütles neiu tasa*). (Vt Eslon 2017b: 36–38)

Kirjeldatud mustri komponentide leksikaalsemantilise, morfosüntaktilise ja funktsionaalse varieerumise seaduspärasusi pole kabinetivaikuses välja mõeldud ega muul moel leiutatud. Muster ja selle

komponentide vahelised seosed on avastatud tekstikorpusest andmekaevele tugineva lingvistilise klasteranalüüsi abil, mille tulemusel saab uudseid andmeid keele elementide süntagmaatilise kombineerimise reeglite ja funktsionaalse võimekuse kohta. Konstruktsioonid seda ei kajasta, nendes tulevad esile keele elementide abstraktsed kategoriaalsed seosed. Samas on mustri põhjal (verb-substantiiv-adverb) üsna lihtne eristada analüütilisi ja sünteetilisi verbe: kõne- ja tuumverbide leksikaal-semantilisse rühma kuuluv põhiverb on reeglina indikatiivi imperfekti ainsuse 3. pöördes, kuid analüütiliste verbide moodustamisel järgnevad sellele substantiiv genitiivse totaalobjekti funktsioonis ja partikkel (*pööras pilgu ära*-tüüpi muster), samas kui sünteetilise verbi vormile järgnevad nominatiivne subjekt ja määrus (*küsis Tehvan pahaselt*-tüüpi muster). Siit tulenevad nii reeglid kui ka skeem, mille alusel saab analoogia põhjal õppida kasutama analüütilisi ja sünteetilisi verbe. Muster kuulub **aktiivse grammatika**, kõne- ja tuumverbid **aktiivse sõnavara** alla.

Niisiis toovad keelekasutusmustrid välja lekseemide tüüpilised semantilis-süntaktilised ja funktsionaalsed seosed, mille tunnetamise ning taasloomise alusel hakkab õppija järk-järgult mõistma tekstide sisu, tähenduste kujunemist, sõnavormide kooskasutust, omandab aktiivsed keelekasutusmustrid, mis väljendavad nt modaalseid ja ekspressiivseid hinnanguid või on vajalikud terviklike sidusate tekstide produtseerimisel. Erinevat liiki konstruktsioonide põhjal niisuguseid andmeid ei saa, sest fookuses on sõnavormide abstraktsed kategoriaalsed ja semantilis-süntaktilised seosed, mitte aktiivsed süntagmaatilised seosed neile omaste vormi- ja sõnakasutuse piirangute ning diferentseeritud tekstifunktsioonidega.

Järgnevalt lühiülevaade uurimustest, milles on kasutatud lingvistilist klasteranalüüsi.

3. Teksti keelekasutusmustrite uurimisest

Lingvistilist klasteranalüüsi on tänaseks kasutatud keelemustrite diakroonsel (nt Trainis 2017) ja sünkroonsel (nt Trainis & Allkivi 2014;

Eslon 2014a; 2014b; 2017b; Paeoja 2015; Eslon & Paeoja 2015) kirjeldamisel, keelevariantide võrdlemisel (nt Voolaid 2018; Allkivi 2016a; 2016b; Eslon 2013), keeleoskustasemete lingvistilise sisu avamisel (nt Voolaid 2018; Allkivi 2016a), tasemespetsiifiliste ja õppija keelekasutust tervikuna iseloomustavate mustrite eristamisel ning keeleoskustaseme automaatsel hindamisel (nt Hallik 2015; Kossinski 2018), sidususe leidmisel õpiku tekstide temaatilise ja üldsõnavara ning grammatikateemade vahel (Ševtšenko 2014; Tšernõšuk 2016), pedagoogilistel eesmärkidel tervikuna (Eslon 2017b).

3.1. Diakroonne aspekt

Jekaterina Trainis (2017) on võrrelnud 1890. ja 1990. aastate ilukirjanduskeele kasutust ning toonud esile sajandi vältel aset leidnud statistiliselt olulised nihked: a) kvantitatiivsed, nt sama mustri leksikaalne varieerumine on sajandi vahetudes muutunud varasemast tunduvalt laiemaks ja leksikaalsemantiliselt mitmekesisemaks (iseloomulik adverbi, verbi ja adjektiivi kasutusele); b) kvalitatiivsed, nt sama mustri komponentidel on võrreldud perioodidel erinev morfosüntaks ja funktsioonid, mis viitab lausestuse muutumisele. Nii on 1890. aastatel mustri adverb + substantiiv + substantiiv komponentide sagedamad funktsioonid ajamäärus + eestäiend ainsuse genitiivis + objekt ainsuse partitiivis, nt *jälle seaduse võimu*. Sajand hiljem on selle mustri kasutuses toimunud kvalitatiivne nihe: viimane substantiiv on kinnistunud ainsuse nominatiivis subjekti funktsioonis, nt *vahest armastuse puudus*, *lausa kulla väärtus*.

1890. ja 1990. aastate ilukirjanduskeele mustrite võrdlusandmed kinnitavad eesti keelele omast üldist arengutendentsi suurema analüütilisuse poole. Eriti selgelt tuleb see esile adverbiga. 1890. aastatel ebaolulistest adverbi sisaldavatest mustritest on sajandi möödudes saanud kinnistunud morfosüntaksi ja tekstifunktsioonidega mustrid, milles esineb uusi ühendverbe ja liitseid adverbilisi üksuseid, nt *eest ära*, *sugugi enam*, *ometi ka*, *vist küll*. Suurenenud on ka konjunktsiooni- ja adjektiiviga algavate mustrite osakaal, mis paistab olema seotud

varasemast keerulisema lausestruktuuri ja laieneva epiteetide kasutamisega. Adpositsioonialguliste mustrite valik ja adpositsioonide hulk on varasemaga võrreldes kahanenud ning näitab kinnistumise tendentsi.

3.2. Sünkroonne aspekt

Kirjeldatud on 1990. aastate ilukirjandustekstide keelekasutusmustreid (Trainis & Allkivi 2014), lähemalt analüüsitud verbi vasak- ja paremkonteksti (Eslon 2017a; 2014a; 2014b), otsitud sarnasust samatähenduslike sünteetiliste ja analüütiliste verbipaaride nagu *töötama – tööd tegema* kasutuses, kuid leitud vaid morfosüntaktilisi piiranguid, mis diferentseerivad nende tähenduskomponente ja rektsioonistruktuure (Paeoja 2015; Eslon & Paeoja 2015).

Uurimused on näidanud, et mustrid verbist paremal ning vasakul liigituvad adverbi sisaldavateks ja adverbita järjenditeks. Esimest liiki mustreid kasutatakse rohkem teksti sidususvahenditena, samuti pragmaatiliste ja subjektiivmodaalsete hinnangute väljendamiseks. Neis tulevad esile analüütilised üksused nagu liitsed adverbikooslused ning analüütilised sõnad, konstruktsioonid ja sidususvahendid. Teist liiki mustrid on reeglina semantiliselt ja morfosüntaktiliselt kinnistunud üksused (sh kollokatsioonid ning idioomid). Kuigi nende osakaal jääb esimest liiki järjenditele kõvasti alla, pole nad tekstiloomes sugugi vähem olulised.

Adverbi olemasolu/puudumine mustris on aluseks ka sagedatele diskursuspõhistele sõnajärjemallidele: a) kaks järjestikust adverbi raamistavad verbi (*alles täna tuli selgesti esile*); b) adverb raamistab liitpredikaati või jaatava ja eitava kõne liitseid verbivorme (*ammu oleks pidanud kirjutama järjekindlalt*); c) substantiiv ja adverb raamistavad verbi finiitvormi (*eile hommikul sulges uksed lõplikult*); d) kaks järjestikust substantiivi määrusliku täiendi või genitiivatribuudi funktsioonis raamistavad verbi (*sadama tuled värelevad mere pinnal*) jne. (Vt Eslon 2017a; 2014a; 2014b; 2013)

3.3. Keelevariantide ja keeleoskustasemete võrdlemine

Lingvistilise klasteranalüüsi tulemustel on avaraid rakendusvõimalusi eesti keele tasemeoskuste lingvistilise sisu esile toomiseks, kirjeldamiseks ja mudeldamiseks.

Kais Allkivi on võrrelnud C1-tasemega vilunud keelekasutaja ja emakeelekõneleja tekstide keelemustreid verbiga algavate ning lõppevate tetragrammide põhjal (vt Allkivi 2016a; 2016b; 2016c). Näiteks emakeelekõnelejate vabas vormis kirjutatud arvamustes on suurim osakaal mustritel, kus verbile järgneb või eelneb vähemalt kaks substantiivi (nt *avastas Marsi pinnalt kraatri*, *kui elanike arv väheneb*), samas kui C1-taseme tekstides eelistatakse pronoomenit ja substantiivi (nt *et see raamat on*, *seob oma tuleviku ainult*). Tegemist on lausestuste erinemisega. C1-taseme sõnastus sarnaneb pigem B1-taseme keelekasutusega, kus pronoomenit sisaldavaid mustreid on tugevalt ülekasutatud, eriti demonstratiivpronoomenit *see* (vt Voolaid 2018). Nende osakaal on C1-tasemele jõudes küll kahanenud, kuid säilitab juhtpositsiooni.

Võrreldes emakeelekõnelejaga on C1-taseme vilunud keelekasutaja tekstides suurem osakaal ka eitaval kõnel, predikatiivil ja täistähenduslikul *olema*-verbil. Harvem kasutatakse mineviku liitajavorme perfekti ja imperfekti, infiniitseid verbivorme ning subjekti ja predikaadi pöördjärge. Allkivi on neid nähtusi käsitlenud kui lingvistilisi markereid, mis eristavad C1-taseme õppija keelekasutust emakeelekõnelejast. Niisuguste markerite hulka kuuluvad ka noomeni käändevormid, modaalverbide ja adverbide esinemus ning funktsioonid jm. Nende olulisust keeleoskustasemete hindamisel on mõõdetud χ^2-testi põhjal ning saadud kinnitust, kas erinevus võrreldavate nähtuste vahel on eri tasemetel juhuslik või tegelik.

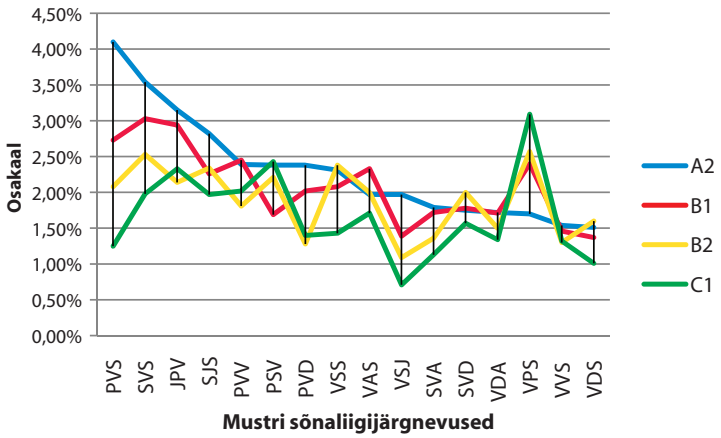
Katrin Voolaid (2018) on analüüsinud B1-taseme soome ja vene lähtekeelega õppijate tekstide keelemustreid ning võrrelnud neid sihtkeele andmetega, mis võetud allikast Trainis & Allkivi 2014. Selgus, et soome lähtekeelega õppijate keelekasutus on sihtkeelepärasem kui vene

lähtekeelega õppijatel: kahest-kolmest samasse rühma kuuluvast mustrist hakatakse tavaliselt kasutama ühte, ent vene lähtekeelega õppijatel on nii mustrite rühmi kui ka neis sisalduvaid mustreid tunduvalt rohkem, kahe-kolme eelistatud mustri ning ülejäänud mustrite osakaal tekstides jaotub ühtlasemalt. Kas see on õpetamise tulemus või mängivad oma osa ka vahetud keelekontaktid, pole esialgu tõestatud.

Võrdlus sihtkeele andmetega näitab B1-taseme keelekasutusmustrite olulist üle- ja alakasutust. Ülekasutus on märk liiasusest, mis põhjustatud sellest, et eelistatakse kindlat hulka mustreid, mille kasutamises ollakse veendunud (iseloomulik soome lähtekeelega õppijate tekstiloomele). Alakasutus viitab õppija keelekasutuse erinemisele sihtkeelest, mis ei sega suhtlust, kuid emakeelekõneleja tunnetab sõnastuse ebaloomulikkust. B1-taseme keelekasutuse olulisim sedalaadi erinevus tuleb esile verbiga algavate mustrite rühmas, mis sihtkeeles lõpevad tavaliselt adverbiga, ent vene lähtekeelega õppijatel substantiivi ja soome lähtekeelega õppijatel adjektiiviga. Tegemist on verbi paremkonteksti V2-sõnajärje malliga (nt *tuleb ta jälle*), mida vene lähtekeelega õppija väldib, eelistades *ma sain palju kingitusi*-tüüpi järjendit. Soome lähtekeelega õppija ei väldi V2-sõnajärge, kuid see pole eelistatuim muster (esinemuse poolest teisel kohal). Järelikult on tegemist vene lähtekeelega õppija keelekasutuses ilmnenud spetsiifilise kõrvalekaldega, mille põhjused võivad olla nii lingvistilised (V2-sõnajärg kuulub vene keele ekspressiivse süntaksi vahendite hulka) kui ka pedagoogilised.

3.4. Teksti keeleoskustaseme ja õpiku hindamine

EVKK kodulehel töötab veebirakendus, mis annab esmase vastuse sisestatud teksti keeleoskustaseme kohta (arvestab sõna, lause ja teksti pikkusega). Samal eesmärgil on arendaja Virgo Hallik (2016) leidnud lingvistilise klasteranalüüsi abiga A2-, B1-, B2- ja C1-taseme tekstidest 16 kokkulangevat mustrit ning vaadelnud nende esinemuse dünaamikat (vt joonis 1).



Joonis 1. *Oluliste keelekasutusmustrite esinemus*

Keeleoskuse arenedes muutub osa mustreid teistest olulisemaks, ühed kerkivad esile esmakordselt, teised kaovad. Liikudes A2-tasemelt kõrgemale, võib mustrite leksikaalne ja morfosüntaktiline kirje suuremal või vähemal määral varieeruda, olla sarnane ja erineda. See tähendab, et mustri komponentide sõnaliik ei muutu, küll aga varieeruvad vormid ja nende funktsioonid. Mustrite põhjal saab tuvastada õppijakeele tüüpsõnavara, leida iga taseme aktiivse sõnavara ja vormistiku, tuua esile sõnaliikide ja vormide dünaamika, stereotüüpse ning kivinenud kasutuse juhtumid.

Teksti keeleoskustaseme hindamise automaatset rakendust arendanud Janek Kossinski (2018) tegi katse integreerida lingvistilise klasteranalüüsi meetodil saadud tulemusi masinõppe meetoditega. Javas töötavat programmi saab käivitada käsurealt.

Lingvistilist klasteranalüüsi on kasutatud ka põhikooli eesti keele kui teise keele õpikute hindamiseks (Ševtšenko 2014; Tšernõšuk 2016). Analüüsitud on 8. ja 9. klassi õpikute tekstide[2] keelekasutusmustreid,

[2] "Eesti keele õpik vene õppekeelega koolile. 8. klass" (Tallinn: ILO, 2004) ning "Eesti keele õpik vene õppekeelega kooli 9. klassile" (Tallinn: Koolibri, 2009).

leitud tekstide aktiivne sõnavara ja grammatika ning hinnatud selle vastavust õpiku grammatikateemadega ja temaatilise sõnavara arendamisega. Saadud tulemused näitavad üheselt, et õppekavaga ette nähtud teemade läbimise, tekstide aktiivse sõnavara ja grammatika ning nende omandamiseks mõeldud harjutuste keelekasutuse vahel puudub igasugune seos. Õppetekstid peaksid järk-järgult laiendama temaatilist sõnavara, tegelikult aga korratakse klassist klassi üldsõnavara, sedagi üsna piiratud mahus. Grammatikateemad ei haaku omavahel ega teksti keelekasutusmustrite morfosüntaksiga. 9. klassis käsitletavatel teemadel, tekstide sõnavaral ja grammatikal puudub loogiline, sisuline ja didaktiline seos sellega, mida õpiti 8. klassis jne. Vajadus uute kontseptuaalselt sidusate õpikute järele on ilmne. Tuleb kasuks, kui koostamisel lähtutaks keelekasutusmustrite esinemusest, nende leksikaalsemantilise ja morfosüntaktilise varieerumise keerukusest (vt ka Eslon 2017a: 230–234).

4. Probleemidest

Lingvistilist klasteranalüüsi saab rakendada mitte ainult keelealastes, vaid ka teistes valdkondlikes uurimustes, kus tegeldakse loomuliku keele tekstidega, olgugi et uurimisobjektid ja eesmärgid on seejuures erinevad. Samal meetodil saadud objektipõhiste ning omavahel võrreldavate tulemuste klassifitseerimine, analüüs ja kirjeldamine ühtse universaalse skeemi alusel võimaldab integreerida erinevaid teadmisi, täiendades rööbiti ka teaduskeele mõistevara. Lisaks uute mõistete kujunemisele toimub vanade ümbermõtestamine. Juba kasutuses üldmõisted rändavad ühest valkonnast ja teooriast teise. Neid on nimetatud **rändavateks** ehk **rändmõisteteks** (vt Tamm 2011: 22). Selles on oma metodoloogiline iva, sest integreeritud teadmiste loogika ühendab analoogia põhjal erinevates valdkondlikes uurimustes kasutatud mõisteid, avardades nii ka teoreetilise mõtte liikumist. Mida ühtedes mõisteseostes näib võimatu kirjeldada, see leiab loomuliku ja täpsema sõnastuse teistes mõisteseostes. Järelikult pole tarvis rändmõistetest loobuda, vaid analoogia põhjal need üle võtta ning põhjendada, miks see on õigustatud.

1) Seletades Klastrileidja toimesüsteemi, oleme kasutanud mõisteid **klaster** ehk rühm sarnaseid objekte (antud juhul sarnase vormistusega **n-gramme**) ning **klasteranalüüs** ehk **klasterdamine**, st sarnaste elementide rühmadesse jaotamine ehk statistiline klassifitseerimistehnika üldistavate tüpoloogiate tarvis (vt Law 2007: 332) või klassifitseerimissüsteemide leidmiseks (Gore 2000: 300).

Traditsiooniliselt on klasteranalüüsi seostatud peamiselt kolme lähenemisviisiga: a) hierarhiline klasterdamine (kasutatakse juhul, kui objekte on suhteliselt vähe või klastrid erinevad üksteisest suhteliselt selgelt); b) k-keskmiste klasterdamine (kasutatakse siis, kui objekte on palju; uurija määrab klastrite arvu) ja c) kombineeritud meetodid (andmete järkjärguline jaotamine, kuni igas rühmas on ainult üks objekt). (Vt Everitt 1997: 466, 468–472; Remm jt 2012: 75–77; Mooi & Sarstedt 2011: 241) Lingvistiline klasteranalüüs ei ühti ühegagi loetletud kolmest lähenemisest, sest analüüsi käigus ei arvutata objektide kaugusi üksteisest, kuid samas on kõigil aluseks sama põhimõte – andmekaeve. See, et andmekaevet saab rakendada erinevate algoritmide põhjal, ei muuda asja olemust, kuid seletab meie puhul rändmõistete **klaster** ja **klasteranalüüs** kasutamist. Põhjendamist vajavad erinevad andmekaeve tehnikad, mistõttu oleme sisse viinud täpsustuse, nimetades Klastrileidjaga tehtud klasterdamist **lingvistiliseks klasteranalüüsiks** (klasterdamine toimub süntaksianalüsaatori märgendite põhjal). Korpuslingvistikas räägitakse ka **morfosüntaktilisest klasterdamisest**, mille käigus rühmitatakse kõik tekstis leiduvad sõnavormid formaalsete morfosüntaktiliste tunnuste alusel, nt koondades ühte kõik verbi finiitvormid, mida on kasutatud kindla kõneviisi oleviku ainsuse 3. pöördes, või kõik mitmuse nimetavas käändes kasutatud substantiivid (vt Sirts jt 2014; Sirts 2015), samuti **semantilisest klasterdamisest**, mille käigus otsitakse lähedase tähendusega sõnu ja jagatakse need kontekstipõhise jaotumise alusel leksikaalsemantilistesse rühmadesse, nt kõik tekstis esinevad modaal-, liikumis-, kõne-, emotsiooni- jm verbid (vt Lagus & Airola 2005).

Terminoloogilist täpsustust vajab ka **n-grammi** mõiste. Meie arusaam n-grammist lingvistilises klasteranalüüsis on analoogne Wiersma

jt (2011) ning Ivaska (2015) käsitlusega, kus eristatakse vormilt ja/või funktsioonilt sarnaseid järjendeid, mille esiletoomine põhineb samuti morfo- ja süntaksimärgenditel. Leitud n-grammid väljastab Klastrileidja nii morfosüntaktiliste märgendite järgnevusena kui ka sellele vastavate keelenäidetena, mis reastatud sageduse alusel. Lingvistika seisukohalt on tegemist sõnavormide statistiliste kooskasutusmustritega tekstis (nt *nii selgesti meeles*, *ennevanasti on olnud*, *seejärel silmitseb mõtlikult*, *veel kaua-kaua tema*, *hirmus ilusad valged*, *nägupidi omad ja*, *otse silmade ees*, *möödus juba seitsesada*), mida loomulikus keelekasutuses esineb suuremal või vähemal määral ning mida tajutakse erineva vormistuse ja funktsionaalse võimekusega tekstiliste terviküksustena. Emakeelekõneleja suudab kerge vaevaga taastada niisuguseid mustreid ümbritseva konteksti (nt <mis? on/oli> *nii selgesti meeles*, <mis? seisis> *veel kaua-kaua tema* <silmade ees>), kuigi tegemist on statistiliselt eristunud järgnevustega, mitte lingvistiliste konstruktsioonidega, mitmesõnaliste leksikaalsete üksuste või kimpudega, ka mitte väljendite, püsiühendite ja idioomidega. Seepärast ongi mõttekas jätta kasutusse üldistav mõiste **n-gramm**, mille pikkust täpsustavad **uni-, bi-, tri- ja tetragramm**.

2) Teine küsimuste ring on seotud rändmõistega **hierarhia**, mille all mõistetakse tavaliselt objektide süstematiseerimis- või jaotamispõhimõtet alt-üles või ülalt-alla printsiibil ehk suunas konkreetsemalt abstraktsemale ja vastupidi. Probleem seisneb selles, missuguste objektide hierarhiaid luuakse ning mille põhjal. Lingvistilise klasteranalüüsi puhul oleme küsimuse lahendamisel lähtunud jällegi Klastrileidja toimesüsteemist, mis rühmitab sarnase vormistusega n-grammid klastritesse. Edaspidine klastrite ühendamine alamklassideks ja viimaste ühendamine klassideks on selle loomulik jätk, mille tulemusel lisandub hierarhiasse veel kaks üldistuse tasandit. Tegemist ei ole keele elementide lingvistilise hierarhiaga ega ka keelesüsteemi lingvistiliste tasandite hierarhiaga, vaid lingvistilise klasteranalüüsi käigus objektipõhiselt leitud mustrite hierarhiaga, millel on klassifitseeriv-süstematiseeriv tähendus. Selles toimimisviisis leidub analoogiat dendrogrammi ehk hierarhilise klasterdamise tulemust visualiseeriva struktuuri tekitamisega, kuid nii

lingvistilise klasteranalüüsi objekt kui ka andmete nominaaltunnused erinevad dendrogrammi aluseks olevatest objektidest ja tunnustest. Seetõttu on mõttekas jääda **lingvistilise klasteranalüüsi hierarhia** mõiste juurde.

3) Kolmas küsimus, mille oleme samuti lahendanud lingvistilise klasteranalüüsi toimesüsteemist lähtudes, seondub üldkasutatava rändmõistega **struktuur**. Selle all mõistame mustri komponentide sõnaliigi järgnevusi, mis on mustrite eristamise universaalne lingvistiline tunnus. Näiteks mustri **_V_** aux neg @NEG + **_V_** main indic pres ps neg #FinV #Intr @+FMV + **_D_** @ADVL (*ei tule enam*) struktuuri tähistav lühend on VVD (verb + verb + adverb). Erinevate mustrite struktuurid on aluseks nende rühmitamisel ja lingvistilisel tõlgendamisel n-grammide, klastrite, alamklasside ning klasside tasandil. Kuna sõnaliigid klassifitseeruvad lingvistikas morfoloogia alla, siis saab kõiki sõnaliigijärgnevusi nimetada analoogia põhjal **morfoloogilisteks struktuurideks**, kuid loomulikult ei tähista see mõiste morfoloogiastruktuure lingvistikas, vaid keelekasutusmustri komponentide sõnaliigilist järgnevust. Tegemist on klassifitseeriva rändmõistega, mis pole välja mõeldud, vaid tuleneb klasteranalüüsi toimesüsteemist.

5. Kokkuvõttev arutelu

Viimaste kümnendite humanitaaria valdkonna arengus on tähelepanu koondunud peamiselt kolmele küsimusele: interdistsiplinaarsus, metodoloogilised pöörded ja rändavad mõisted (vt Tamm 2011: 16 jj). Lingvistilise klasteranalüüsi meetodil tehtud uurimused paigutuvad sellesse raamistikku, võimaldades analüüsida ja kirjeldada keele elementide süntagmaatilisi kombineerumisreegleid metodoloogiliselt uutel alustel. Seetõttu saame keelekasutuse kohta uudseid andmeid, mis on väärtuslikud nii rakenduslikus kui ka teoreetilises plaanis. Eespool tõime konkreetseid näiteid.

Keeleteaduslike uurimuste fookus on ammu nihkunud keelelt kui lingvistika objektilt selle uurimiseks sobilikele tehnikatele, st

konkreetsetelt teooriatelt algoritmipõhisele statistilisele või statistiliskombinatoorsele analüüsile, milles sõnade ja vormide asemel opereeritakse arvandmete ning formaalsete tunnustega. Näiteks eelmise sajandi keskel köitis tähelepanu Morris Swadeshi leksikaal-statistiline teooria, mille kohaselt koosneb põhisõnavara universaalseid üldmõisteid tähistavatest sõnadest, mille koosseis on muutumatu ja erinevates keeltes 81%–86% ulatuses kokkulangev. (Vt Klimov 1961: 241) Statistiliste ja kombinatoorsete tehnikate rakendamine loomuliku keele analüüsimisel ning kirjeldamisel on aluseks tarkvara arendamisele, eesti keele reeglipõhist süntaksianalüsaatorit on kombineeritud statistikaga (Muischnek jt 2012), teksti keerukuse uuringutest on välja kasvanud eesti keele sagedussõnastik jne.

Loengutes keele ning mõtlemise vahekorrast ennustas Noam Chomsky eriti suurt tulevikku just matemaatilisele ehk arvutilingvistikale, mille areng pole seotud andmete loendamise, vaid abstraktsete printsiipide ja struktuuride leidmisega, mis määravad inimese mõteprotsesside organiseerumist keeleteadmistes (Chomsky 1972: 90). Üks paremaid sellekohaseid näiteid on Meľčuki-Zholkovsky (1984) seletavkombinatoorne sõnastik, kus lekseemi kirjeldatakse seostes teiste lekseemidega tähenduskomponentide kaupa sama formaalse skeemi alusel. Igor Meľčuki (1995: 81–133) arvates lubab formaalsetele tunnustele rajanev universaalne skeem võrrelda erinevate lekseemide semantilist ja grammatilist valentsi ning esile tuua lekseemide süntagmaatilisi kombineerimisreegleid, millest omakorda tulenevad edasised teoreetilised lahendused lingvistikas, keeletehnoloogias, leksikograafiliste rakenduste arendamisel jne. Tänaseks on see kinnitust leidnud nagu ka tõsiasi, et keele elementide statistiline tekstiline jaotumine kujutab endast samasugust universaalset skeemi, mille põhjal on avastatud varjatud semantilisi klassifikatsioone (Šajkevitš 1976: 360; Šajkevitš jt 2013: 15–18).

Semantilised seosed ilmnevad keele elementide formaalses tekstilises jaotumises, piiravad elementide kombineerimisvõimalusi, tingivad leksikaalsemantilisi ja morfosüntaktilisi valikuid, mida konstruktsiooni tasandi kvalitatiivne lingvistiline analüüs adekvaatselt ei kajasta. Siit ka

põhjus, miks grammatikareeglitest ja konstruktsioonidest pole keele õppimisel suuremat kasu, kui pole just spetsiifilist teaduslikku eesmärki või huvi grammatikasüsteemide iseärasuste vastu. Küll aga on kasu lingvistilise klasteranalüüsi abil tekstist eristatud keelekasutusmustritest, kui eesmärk on õppida võõras keeles suuliselt suhtlema ja tekste kirjutama. Siis läheb vaja keele elementide järjestikuse kooskasutuse reegleid, mis peaksid olema õppija keeleteadmiste alus. Chomsky järgi on nt fonoloogiauuringud üha selgemalt tõestanud, et häälikusüsteemide olemus ei peitu nende kirjeldamiseks kasutatud struktuurides ja mudelites (konstruktsioonides ja grammatikareeglites – autorite lisandus), vaid keerulises reeglistikus, mille alusel neid süsteeme ja mudeleid saab üles ehitada, modifitseerida ning üksikasjalikumaks muuta. Sedalaadi reeglistik rajaneb universaalsetel abstraktsetel tunnustel, mis omavahel keerulisel moel põimunud, võimaldades keelestruktuure matemaatiliselt käsitleda. (Chomsky 1972: 93) Sama on ammu tõestanud keelestruktuuride kombinatoorse statistilise analüüsi tulemuste rakendamine masintõlkes, aluskeele rekonstrueerimisel (Andrejev 1965) ning keele elementide süntagmaatiliste kombineerimisreeglite leidmisel (Andrejev 1967), sh ka eesti allkeelte statistilis-kombinatoorse morfoloogiamudeli loomisel (Holm 1965).

Eespool tõime näiteid ja seletasime Klastrileidja toimesüsteemi: mida programm tekstis eristab, sageduse põhjal esile toob ja rühmitab. Selgus, et erinevates mustrites on komponentide vahel kindlad kooskasutuse piirangud (reeglid), mida saab lingvistiliselt tõlgendada ja sõnastada. Kuigi nii andmekaeve meetod kui ka lingvistiline klasteranalüüs kuuluvad tehnoloogia valdkonda, tugineb Klastrileidja toimesüsteem formaalsetele lingvistilistele tunnustele ja nende järjestikuse kooskasutuse arvandmetele. Tegemist on mitte ainult omaette meetodi, vaid ka uurimissuunaga, mis annab objektipõhiseid mõõdetavaid tulemusi, toob ühtse universaalse skeemi alusel (sõnaliikide süntagmaatiline järgnevus tekstis ehk mustri struktuur) esile loomulikke lingvistilisi klassifikatsioone, mida saab kirjeldada kui hierarhiat, liikudes kas üksikult üldisemale (n-gramm > klaster > alamklass > klass) või vastupidi.

Ollakse arvamusel, et statistiliste ja kombinatoorsete tehnikate integreerimine lingvistilise analüüsiga on võimaldanud esile tuua uudseid mõisteseoseid ja keelenähtusi ning kujundanud arusaamu inimkeele tegelikust toimisest tänu suuremahulistele tekstikorpustele (vt Biber jt 2006: 55–58; Gries & Stefanovitsch 2006). Ka lingvistilise klasteranalüüsi meetodil saadud tulemused on uudsed ja väärivad tähelepanu nii teoreetilises kui ka rakenduslikus plaanis, ehkki korpusvalimid pole olnud kuigi mahukad (nt 1990. aastate ilukirjanduskorpus sisaldab ligi 200 000 sõnet, B1-taseme õppijakeele korpus u 120 000 ja C1-taseme korpus u 75 000 sõnet). Vaja on tagada erineva suurusega valimite representatiivsus, andmete statistiline olulisus ja võrreldavus ning tõestada, et esile tulnud sarnasused/erinevused pole juhuslikud.

Lekseemide loomulik järjestus võimaldab ühendada nende sõnastikusemantika abstraktsema süntaktilise semantikaga ning seletada sõnakasutuse tekstilisi (kommunikatiivseid ja pragmaatilisi) funktsioone objektilähedaselt. Langeb ära vajadus küsida näiteks, kas subjekti mõiste taga on loogiline, semantiline või pragmaatiline kategooria, kas subjekti tuleks vaadelda seoses predikaadi argumendistruktuuri või verbi aktandistruktuuriga jne. Aktiivsed seosed subjekti semantiliste rollide ja verbilekseemi semantilise valentsi vahel tulevad ilmsiks tekstikasutuses. Analoogseid semantilisi pesasid võib leida ja süstematiseerida ka formaalsete ja konstrueeritud struktuuride alusel, eristades komponentide semantilisi rolle – agent, objekt, instrument jne, kuid sel on rohkem teoreetiline kui rakenduslik väärtus. Keeleõpe vajab elementide kasutusreeglid, mitte abstraktseid süntaksistruktuure ja grammatikareegleid.

Kirjandus

Allkivi, Kais 2016a. C1-tasemega eesti keele õppijate kirjalik keelekasutus võrdluses emakeelekõnelejatega: samalaadsusi ja nihkeid verbist paremal paiknevas kontekstis [‘Written language use of C1 learners of Estonian and native speakers in comparison: Similarities and differences in verb-initial fourgrams’]. Magistritöö. Tallinn: Tallinna Ülikool. http://www.etera.ee/zoom/20076/view?page=1&p=separate&view=0,0,2481,3508 (30.9.2017).

Allkivi, Kais 2016b. C1-tasemega eesti keele õppijate ja emakeelekõnelejate kirjaliku keelekasutuse võrdlus verbialguliste tetragrammide näitel [‘Written language use of C1 learners of Estonian and native speakers in comparison: Analysis of verb-initial fourgrams’]. – Lähivõrdlusi. Lähivertailuja 26, 54–83. https://doi.org/10.5128/LV26.02

Allkivi, Kais 2016c. Verbist paremal ja vasakul paiknev kontekst C1-tasemega eesti keele õppijate ja emakeelekõnelejate kirjutistes [‘Written language use of C1 learners of Estonian and native speakers in comparison: Analysis of verb-initial and verb-ending fourgrams’]. – XII muutuva keele päev. Ettekannete teesid. Tartu: Tartu Ülikool, 18–19. https://www.keel.ut.ee/sites/default/files/www_ut/mkp_2016_teesid.pdf (12.8.2018).

Andrejev 1965 = Андреев, Николай Дмитриевич 1965. Статистико-комбинаторное моделирование языков [‘Statistical-combinatorial modeling of languages’]. Сборник научных статей. Отв. ред. Николай Дмитриевич Андреев. Москва, Ленинград: Наука.

Andrejev 1967 = Андреев, Николай Дмитриевич 1967. Статистико-комбинаторный метод в теоретическом и прикладном языкознании [‘Statistical-combinatorial method in theoretical and applied linguistics’]. Ленинград: Наука.

Biber, Douglas, Susan Conrad, Randi Reppen 2006 [1998]. Corpus Linguistics. Investigating Language Structure and Use. New York: Cambridge University Press.

Chomsky 1972 = Хомский, Ноам 1972. Язык и мышление [‘Language and Mind’]. Перевод с английского Б. Ю. Городецкого. Москва: Изд-во Московского университета.

Eslon, Pille 2013. Kahe keelekasutusvariandi võrdlus: morfoloogilised klassid ja klastrid [‘The comparative study of language use: morphological classes and clusters’]. – Lähivõrdlusi. Lähivertailuja 23, 13–38. https://doi.org/10.5128/LV23.01

Eslon, Pille 2014a. Adverbi sisaldavate struktuuride tekstifunktsioonidest eesti ilukirjandus- ja õppijakeeles [‘On the textual functions of adverbial structures in literary Estonian and Estonian learner language’]. – Lähivõrdlusi. Lähivertailuja 24, 15–46. https://doi.org/10.5128/LV24.01

Eslon, Pille 2014b. Morfosüntaktilise ja leksikaalse varieerumise piiridest: ilukirjandus- ja õppijakeele kasutusmustrite võrdlus [‘Constraints on morphosyntactic and lexical variability’]. – Eesti Rakenduslingvistika Ühingu aastaraamat 10, 55–71. https://doi.org/10.5128/ERYa10.04

Eslon, Pille 2017a. Kasutuspõhise keelekäsitluse pedagoogiline perspektiiv ['Usage-based language description: Linguistic cluster analysis ant it's perspectives for pedagigical purposes']. – Mäetagused 69, 217–242. https://doi.org/10.7592/MT2017.69.eslon

Eslon, Pille 2017b. Keelekasutusmustrid verbist paremal: morfosüntaktiline ja leksikaalsemantiline varieerumine ['Patterns of language use found on the right periphery of the verb: Morphosyntactic and lexico-semantic variability']. – Lähivõrdlusi. Lähivertailuja 27, 17–64. http://dx.doi.org/10.5128/LV27.01

Eslon, Pille, Heleriin Paeoja 2015. Samatähenduslike sünteetiliste ja analüütiliste verbide kasutamine ['Use of the synonymous synthetic and analytical verbs']. – Lähivõrdlusi. Lähivertailuja 25, 63–104. https://doi.org/10.5128/LV25.04

Everitt, Brian S. 1997. Cluster analysis. – John P. Keeves (Ed.), Educational Research, Methodology, and Measurement: An International Handbook. Australia: Flinders University of South Australia, 466–472.

EVKK = Eesti vahekeele korpus. http://evkk.tlu.ee (12.3.2018).

Gore, Paul A. Jr. 2000. Cluster analysis. – Howard E. A. Tinsley, Steven D. Brown (Eds.), Handbook of Applied Multivariate Statistics and Mathematical Modeling. New York: Academic Press, 297–321. https://www.hse.ru/data/2012/01/04/1262163878/Gore%20Paul%20A.%20(2000)%20Cluster%20Analysis.pdf (18.9.2017).

Gries, Stefan Th., Anatol Stefanovitsch (Eds.) 2006. Corpora in Cognitive Linguistics. Corpus-Based Approaches to Syntax and Lexis. Berlin, New York: Mouton de Gruyter.

Hallik, Virgo 2016. Eesti vahekeele korpuse klasteranalüüsi vahendite kasutamine teksti keeletaseme prognoosimisel ['Using Estonian interlanguage corpus cluster analysis tools to predict language level of text']. Bakalaureusetöö. Tallinna Ülikool, digitehnoloogiate instituut.

Holm 1965 = Хольм, Хелье Х. 1965. Выделение парадигмы первого морфологического типа на различных подъязыках при статистико-комбинаторном моделировании эстонской морфологии ['Identification of the paradigm of the first morphological type in various sublanguages for statistical-combinatorial modeling of Estonian morphology']. Ученые записки Тартуского государственного университета 172. Тарту.

Ivaska, Ilmari 2015. Edistyneen oppijansuomen konstruktiopiirteitä korpusvetoisesti: avainrakenneanalyysi ['Corpus-driven approach towards

constructional features of advanced learner Finnish: Key structure analysis']. Väitöskirja. Turun yliopiston julkaisuja C-409. Turku: Turun yliopisto.

Kitsnik, Mare 2018. Iga asi omal ajal: eesti keele B1- ja B2-taseme verbikonstruktsioonid keeleoskuse arengu näitajana ['All in good time: Estonian B1- and B2-level verbal constructions as indicators of the development of language proficiency']. Tallinna Ülikool. Humanitaarteaduste dissertatsioonid 43. Tallinn: Tallinna Ülikool. https://www.etera.ee/zoom/41182/view?page=1&p=separate&view=0,0,2067,2834 (11.8.2018)

Klimov 1961 = Климов, Георгий Андреевич 1961. О лексико-статистической теории М. Сведоша ['About the lexico-statistic theory by M. Swadesh']. – Вопросы теории языка в современной зарубежной лингвистике. Москва: Изд-во Академии Наук СССР, 239–253.

Kossinski, Janek 2018. Masinõppel rajaneva tarkvararakenduse loomine keeleoskustaseme ennustamiseks ['Development of a language skill prediction software using machine learning']. Bakalaureusetöö. Tallinna Ülikooli digitehnoloogiate instituut.

Lagus, Krista, Anu Airola 2005. Semantic clustering of verbs – analysis of morphosyntactic contexts using the SOM algorithm. – Acquisition and Representation of Word Meaning: Theoretical and Computational Perspectives. Linguistica Computazionale XXII-XXIII. Pisa-Roma: IEPI, 263–287. https://pdfs.semanticscholar.org/3947/53bcca76302f23ad8ffabe4e91272a2d03a6.pdf (18.1.2018).

Law, Nancy 2007. Comparing pedagogical innovations. – Mark Bray, Bob Adamson, Mark Mason (Eds.), Comparative Education Research. Approaches and Methods. Hong Kong: The University of Hong Kong, 333–364.

Matsak, Erika, Pille Eslon, Jaagup Kippar 2010. Eesti keele sõnajärje vealeidja prototüübi arendamine ["The development of the prototype for an automatic word order error detector for the Estonian language']. – Pille Eslon, Katre Õim (Toim.), Korpusuuringute metodoloogia ja märgendamise probleemid. Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 12. Tallinn: Tallinna Ülikooli Kirjastus, 59–100.

Meľčuk, Igor 1995 [1979]. Semantics of two emotion verbs in Russian: *bojat'sja* '[to] be afraid' and *nadejat'sja* '[to] hope'.– Игорь Мельчук, Русский язык в модели «смысл <=> текст». Москва, Вена: Языки русской культуры, 81–133.

Meľčuk, Igor, Aleksandr Zholkovsky 1984. Explanatory combinatorial dictionary of modern Russian. Wiener slawistischer Almanach 14 (Sonderband). Peter Lang.

Metslang, Helena, Erika Matsak 2010. Kesksete lausekomponentide järjestus õppijakeeles: arvutianalüüsi katse ['Automatic word order analysis of Estonian as a second language: The nuclear sentence']. – Eesti Rakenduslingvistika Ühingu aastaraamat 6, 175–193.

Mooi, Erik, Marko Sarstedt 2011. A Concise Guide to Market Research. The Process, Data, and Methods Using IBM SPSS Statistics. Springer. Chapter 9: Cluster Analysis. Berlin, Heidelberg: Springer, 237–284. https://doi.org/10.1007/978-3-642-12541-6

Muischnek, Kadri, Mark Fišel, Heiki-Jaan Kaalep, Mare Koit, Kaili Müürisep, Heili Orav, Kadri Vare, Haldur Õim 2012. Arvutilingvistika ja keeletehnoloogia Tartu Ülikoolis ['Development of computational linguistics and language technology at the University of Tartu']. – Emakeele Seltsi aastaraamat 57 (2011), 66–102. https://doi.org/10.3176/esa57.05

Muischnek, Kadri, Kaili Müürisepp, Tiina Puolakainen 2016. Estonian Dependency Treebank: from Constraint Grammar tagset to Universal Dependencies. – Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, May 23–28, 2016. ELRA, 1558–1565. http://www.lrec-conf.org/proceedings/lrec2016/pdf/411_Paper.pdf (12.4.2018).

Ots, Sander 2011. Tarkvara statistiliste kollokatsioonide eraldamiseks ning selle rakendus morfosüntaktilises analüüsis ['Software for extracting statistical collocations and its application in morphosyntactic analysis']. Seminaritöö. Tallinna Ülikooli informaatika instituut.

Ots, Sander 2012. Statistikapõhise tarkvara loomine morfoloogiliste kollokatsioonide eraldamiseks eesti keele tekstidest ['Software for morphosyntactic cluster extraction from Estonian texts']. Bakalaureustöö. Tallinna Ülikooli informaatika instituut.

Paeoja, Heleriin 2015. Analüütiliste/sünteetiliste verbipaaride kasutusmustrid 1990ndate aastate eesti ilukirjanduskeeles ['Usage patterns of analytic and synthetic verbal pairs based on 1990s Estonian literature texts']. Magistritöö. Tallinna Ülikooli eesti keele ja kultuuri instituut.

Remm, Kalle, Jaanus Remm, Ants Kaasik 2012. Ruumiliste loodusandmete statistiline analüüs ['Statistical analysis of spatial data']. Õpik-käsiraamat. Tartu: Tartu Ülikooli ökoloogia ja maateaduste instituut.

Sirts, Kairit 2015. Non-parametric Bayesian models for computational morphology. Theses of Tallinn University of Technology C100. Tallinna Tehnikaülikool.

Sirts, Kairit, Jacob Eisenstein, Micha Elsner, Sharon Goldwater 2014. POS induction with distributional and morphological information using a distance-dependent Chinese restaurant process. – Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (2), 265–271.

Šajkevitš 1976 = Шайкевич, Анатолий Янович 1976. Дистрибутивно-статистический анализ в семантике ['Distributive Statistical Analysis of Semantics']. – Принципы и методы семантических исследований. Москва: Наука, 353–378.

Šajkevitš jt 2013 = Шайкевич, Анатолий Я., Владислав М. Андрющенко, Наталья А. Ребецкая 2013. Дистрибутивно-статистический анализ языка русской прозы 1850–1870-х гг. ['Distributive statistical analysis of the language of Russian prose of the 1850s-1870s']. Том 1. Москва: Языки славянской культуры.

Ševtšenko, Marina 2014. Eesti keele kui teise keele 8. klassi õpiku temaatiline sõnavara ja grammatika ['Vocabulary and grammar of the Estonian as a second language textbook (8th class)']. Magistritöö. Tallinna Ülikooli eesti keele ja kultuuri instituut.

Tamm, Marek 2011. Humanitaarteaduste metodoloogia: minevik ja tulevik ['Methodology of humanities: past and future']. – Humanitaarteaduste metodoloogia. Uusi väljavaateid. Tallinn: TLÜ Kirjastus, 9–29.

Trainis, Jekaterina 2015. Linguistic cluster analysis: A method for describing language units and indicating regularities in language. – Wojciech Malec, Marietta Rusinek (Eds.), Within Language, beyond Theories. Vol. III. Discourse Analysis, Pragmatics and Corpus-based Studies. Cambridge Scholars Publishing, 229–243.

Trainis, Jekaterina 2017. Diakroonilised nihked eesti ilukirjanduskeele kasutusmustrites 1890–1990 ['Diachronic shifts in usage patterns of Estonian belletristic language in 1890s–1990s']. – Mäetagused 69, 181–216. https://doi.org/10.7592/MT2017.69.trainis

Trainis, Jekaterina, Kais Allkivi 2014. Ilukirjanduskeelest uue pilguga ['On belletristic language from a new perspective']. – Eesti Rakenduslingvistika Ühingu aastaraamat 10, 283–306. https://doi.org/10.5128/ERYa10.18

Tšernõšuk, Anna 2016. Eesti keele kui teise keele 9. klassi õpiku temaatiline sõnavara ['Vocabulary of the Estonian as a second language textbook (9th class)']. Bakalaureusetöö. Tallinna Ülikooli humanitaarteaduste instituut.

Voolaid, Katrin 2018. Vene ja soome lähtekeelega õppijate eesti keele kasutusmustrid (B1-tase) ['Estonian language usage patterns among Russian and

Finnish students (B1 language proficiency level)']. Magistritöö. Tallinna Ülikooli humanitaarteaduste instituut.

Wiersma, Wybo, John Nerbonne, Timo Lauttamus 2011. Automatically extracting typical syntactic differences from corpora. – Literary and Linguistic Computing 26 (1), 107–124. https://doi.org/10.1093/llc/fqq017

Patterns of language use and linguistic cluster analysis

PILLE ESLON, KAIS ALLKIVI-METSOJA
Tallinn University

For automatic processing of large electronic corpora, different language analysis tools and statistical methods are applied, the choice and combination of which depend on the language, the object and goals of study. In this article, we introduce an integrated software tool Klastrileidja (Cluster Catcher), which has been developed for finding language use patterns, and we give an overview of the study results obtained, using linguistic cluster analysis. The purpose is to explain the possibilities that this method offers for natural language processing, exploring Estonian and learner language use as well as for pedagogical needs.

Keywords: natural language processing; language use patterns; learner language; Estonian

Pille Eslon
Tallinna Ülikooli digitehnoloogiate instituut
Narva mnt 29, 10120 Tallinn, Estonia
peslon@tlu.ee

Kais Allkivi-Metsoja
Tallinna Ülikooli digitehnoloogiate instituut
Narva mnt 29, 10120 Tallinn, Estonia
kais@tlu.ee