doi:10.5128/LV31.01

Eesti keele A2–C1-taseme kirjalike tekstide võrdlev automaatanalüüs

KAIS ALLKIVI-METSOJA
Tallinna Ülikool

Ülevaade. Tänaseni puudub ülevaade eesti keele kui sihtkeele õppijate tekstiloomest eri keeleoskustasemetel, mis põhineks tekstide automaatanalüüsi andmete statistilisel töötlusel. Eesmärk on kindlaks teha, millised arvuliselt mõõdetavad tunnused iseloomustavad A2–C1-taseme eksamite loovkirjutiste leksikaalset keerukust ning sõnaliikide ja -vormide kasutust, olles seejuures keeleoskustasemete piiritlemisel nii statistiliselt kui ka sisuliselt olulised. Esile tulevad järjestikuseid tasemeid (A2–B1, B1–B2, B2–C1) läbivalt ja osaliselt eristavad tunnused. Ühed neist muutuvad lineaarselt kasvavas või kahanevas suunas, teiste tunnuste lõikes ei ole muutused aga samasuunalised ja seostuvad pigem kirjutamisülesande (teksti liik, teema) kui keelekasutuse kasvava keerukusega. Uurimuse tulemused pakuvad uudseid teadmisi keeleõppe seisukohalt ja aitavad edaspidi arendada keeleoskustaseme automaathindajat, tuues välja usaldusväärsemad tunnused tasemete prognoosimiseks.

Võtmesõnad: keele automaattöötlus; keeleoskustasemed; leksikaalne keerukus; morfoloogiline analüüs; kirjalik õppijakeel; eesti keel

1. Uurimuse taust

Euroopa ühtsele keeleoskustasemete skaalale (CEFR 2001, 2018; HTM 2007) vastavate suhtluseesmärkide täitmiseks on keeleõppijal tarvis omandada sihtkeele eriomased leksikaalsed ja grammatilised väljendusvahendid. Samas on keelespetsiifiliste tunnuste kohta üsna vähe empiirilisi andmeid, mis iseloomustavad keelekasutuse edenemist ühelt tasemelt teisele (vt nt Hulstijn 2014; Wisniewski 2017).

Keeleoskustasemete lingvistilise profiili määramiseks pakuvad ainest õppijakeelekorpused, mille kasutamisest sel eesmärgil annab ülevaate Wisniewski (2017). Olulisimad algatused eri tasemete keelekasutuse võrdlevaks uurimiseks pärinevad SLATE-võrgustikult (Second Language Acquisition and Testing in Europe, vt Bartning jt 2010) ja Cambridge'i ülikooli projektist "English Profile" (Harrison & Barker 2015). Omaette uurimissuund on keeleoskustaseme ennustamine masinõppe abil, mis võimaldab anda õppijaile automaatset tagasisidet. Nt saksa-, tšehhi-, rootsi- ja ingliskeelsete tekstide taset on määratud grammatika, sõnavara, teksti sidususe ja vigase keelekasutusega seotud tunnustest lähtudes (Hancke 2013; Rysová jt 2016; 2019; Tack jt 2017; Pilán 2018; Arnold jt 2018; Yannakoudakis jt 2018; Szügyi jt 2019).

Automaatselt on identifitseeritud ka eesti keele õppijate A2–C1-taseme kirjutisi leksikaalsete ja morfoloogiliste tunnuste alusel (Vajjala & Lõo 2014) ning sõnaliigijärjendite põhjal (Kossinski 2018). Vajjala ja Lõo on välja toonud mõningad parima prognoosivõimega tunnused, ent pole selgitanud, kuidas need keeleoskustasemeid eristavad.

Rohkem tähelepanu on osutatud mõningatele tasemeomase keelekasutuse aspektidele, nagu kirjutamisprotsessi sujuvus (Gaitšenja 2019), sõnavara keerukus A2–C1-tasemel (Alp jt 2013), tingiva kõneviisi ja modaaltegusõnade tarvitus B1- ja B2-tasemel (Kitsnik 2018). Kõrvutatud on tegusõnade kasutusmalle C1-taseme eesti keele õppijate ja emakeelekõnelejate arutlevates tekstides (Allkivi 2016; Allkivi-Metsoja 2021), samuti A2- ja B1-taseme soome- ja venekeelsete õppijate keelekasutusele omaseid sõna- ja vormieelistusi (Eslon 2021; Voolaid 2018).

Käimas on keeleoskustasemete sõnavara- ja grammatikapädevuse kirjelduste täpsustamine, mis arvestab mh noorte ja täiskasvanud eesti keele kui sihtkeele õppijate keelekasutusega, kuid peegeldab siiski pigem eri tasemetele seatavaid ootusi (Üksik jt 2021).

Kehtib üldine paradoks, et automaatsele tasemehindamisele keskendunud uurimused ei selgita ega tõlgenda tasemeid eristavate tunnuste dünaamikat keeleoskustasemete lõikes, lingvistilised uurimused aga pühenduvad pigem konkreetsetele keelelistele nähtustele. Ka eesti keeles on siiani vajadus süsteemse kasutuspõhise kirjelduse järele, mis tooks tekstide arvuliselt mõõdetavate ja automaatselt leitavate lingvistiliste tunnuste seast esile keeleoskustasemeid piiritlevad tunnused. Automaatanalüüs võimaldab kommunikatiivsete tasemekirjelduste alusel hinnatud tekste iseloomustada kvantitatiivsest vaatepunktist ning teha üldistusi selle kohta, milliseid keelevahendeid ja mil määral eri tasemega õppijad oma tekstiloomes tegelikult tarvitavad.

Artiklis tutvustatakse osa laiemast uurimusest, mis ühendab keeleoskustaset hindavate ennustusmudelite koostamise ning tasemete erinevust markeerivate lingvistiliste tunnuste kvantitatiivse ja kvalitatiivse kirjelduse, tuginedes seejuures eesti keele A2–C1-taseme eksamite loovkirjutiste automaatse analüüsi andmetele. Selline lähenemine on ka rahvusvaheliselt uudne. Siinse uurimuse eesmärk on teha esmalt kindlaks teksti keerukusega seotud tunnused, mis on nii statistiliselt kui ka sisuliselt olulised järjestikuste keeleoskustasemete eristamisel ja mida võiks seega kasutada sisendina tekstide taseme automaathindaja arendamisel.

Eksamitekste on analüüsitud leksikaalsete ja morfoloogiliste tunnuste alusel (kokku 69 tunnust). Leksikaalsed tunnused seostuvad sõnavara keerukuse eri aspektidega (vt 2.2.1.), morfoloogilised kirjeldavad sõnaliikide ja grammatiliste vormide kasutust. Nende tunnuste leidmiseks tuli tekstid lemmatiseerida (algvormistada) ja morfoloogiliselt märgendada.

Vastust otsitakse järgmistele küsimustele.

1) Milliste lingvistiliste tunnuste lõikes erinevad A2–C1-taseme kirjalikud tekstid ja kuidas?

2) Kuidas ja milliseid A2–C1-taseme tekstide lingvistilisi tunnuseid mõjutab kirjutamisülesanne?

Artiklis kirjeldatakse, missugustele keeleoskustasemetele vastavaid tekste üks või teine tunnus olulisel määral eristab ja kas tunnuste muutumises võib märgata kasvavat, kahanevat või erisuunalist trendi. Vaadeldakse lingvistiliste tunnuste koosesinemise seaduspärasid ja sõltumist eksamiülesandest, sest ka varasemad uurimused on ilmsiks toonud kirjutamisülesande mõju õppijate sõnavara- ja grammatikakasutusele (nt Kuiken & Vedder 2007; Mylläri 2020). Eri tasemete tekstide piiritlemisel saab lugeda usaldusväärsemaks tunnuseid, mille puhul on muutused samasuunalised ja hajuvus tasemete piires võrdlemisi väike.

2. Uurimismeetodid ja materjal

2.1. Tekstivalim

Analüüsiks sobivat valimit ehk korpust koostades tuleb määratleda materjali valikuühik (ingl *sampling unit*) ja üldkogum, mis hõlmab kõiki võimalikke valikuühikuid, antud juhul teatud kriteeriumitele vastavaid kirjalikke tekste. Seejärel määratakse valimiraam (ingl *sampling frame*) – valikuühikute kogum, mille põhjal korpus tegelikult kokku pannakse. Valimiraam on piiritletum kui üldkogum, ent samas võimaldab teha usaldusväärseid järeldusi. (McEnery jt 2006: 19–20) Korpuse representatiivsus rajaneb keelematerjali statistilise üldistatavuse, tasakaalustatuse, aktuaalsuse ja varieeruvuse põhimõtetel (McEnery jt 2006: 13–21).

Siinse uurimuse üldkogumi moodustavad eesti keele kui sihtkeele õppijate tasemeeksamite loovkirjutised, mida on hinnatud vastavaks A2–C1-tasemele. Valimiraam hõlmab tekste, mis on kirjutatud eesti keele tasemeeksamitel 2018. aastal, C1-taseme tekstid ka 2017. aastal. Valimisse kuulub 480 algselt käsikirjalist teksti, mis on saadud SA Innovelt vahemikus 2019. aasta septembrist novembrini – kokku 120 kirjutist iga taseme kohta. Valik on tehtud eksamitöödest, mille kirjutamisosa eest on eksaminand saanud vähemalt 60% võimalikust punktisummast

ehk hinnangu "rahuldav". Tasemeeksamite kirjutamisosa koosneb kahest ülesandest, valimisse on võetud loovtekstid.

Iga keeleoskustaseme eksamikirjutiste seast on tehtud juhuvalik, milleks kasutati R-i paketti Sampling. Et eksamid toimuvad kord kvartalis, siis on võimalusel valitud kõigilt eksamitelt võrdne arv tekste. 2018. aastal sooritati 1608 eksamitööd, mis eeldatavalt kajastavad tasemekohast kirjutamisoskust: A2-tasemel 490, B1-tasemel 711, B2-tasemel 343, C1-tasemel 64 teksti. A2–B2-tasemel valiti 2018. aasta neljast eksamikorrast 30 juhuslikku tööd. Sobivate tööde vähesuse tõttu võeti C1-taseme kirjutiste hulka ka 2017. aasta eksamid (92 teksti). Kuna C1-taseme 156 teksti ei jaotunud 2017.–2018. aasta kaheksa eksami vahel selliselt, et korpusesse saanuks võtta igalt eksamilt võrdse arvu tekste, siis eraldati juhuvalikuga kahe aasta kõigi eksamitööde hulgast 120 kirjutist. Tekstide jaotumine eri eksamikordade vahel suurendab valimi mitmekesisust ja usaldusväärsust, sest nii kirjutamisülesanne kui ka hindajate koosseis erineb eksamiti.

Valimi mahule on piirangu seadnud vajadus eksamitekstid keeleõppija kirjaviisi säilitaval kujul ja pseudonüümitult ümber trükkida, mida on teinud Innove töötaja. Tekstid on talletatud Eesti vahekeele korpuses (EVKK). Lisatud metaandmed (eksaminandi vanus, sugu, haridustase, elukoht ja kodakondsus) ei võimalda isikut tuvastada.

Kirjutamisülesandest tulenevalt varieeruvad eksamitekstide liigid ja teemad: A2-tase – kirja või kuulutuse vormis teated ja kirjeldused (kutse väljasõidule või kodumasinat parandama, automüügikuulutus, viimase reisi kirjeldus); B1-tase – jutustused minevikusündmustest (kohtumine kunagise koolikaaslasega ja viimatine kontserdikülastus), oma soovidest ja harjumustest (lemmikloomade pidamine ja meediakasutus); B2-tase – isiklikud (teemaks tööprobleemi lahendamine või seisma jäänud raamatute äraandmine) ja poolametlikud kirjad (kaebekiri kohalikule omavalitsusele tänavate ja parkide korrashoiu teemal, panus töökoha sünnipäeva puhul ilmuva raamatu koostamisse); C1-tase – arvamusartiklid ja kolleegidele suunatud ülevaated, mis ajendatud osalemisest ühiskondlikult olulist teemat puudutanud üritusel, nt

teabepäeval või seminaril (teemad: sallivus teiste rahvuste vastu, digipädevus, kodanikuühenduste vajadus ja roll, rahva tervis ja töötervishoid, õnnetusjuhtumite ennetamine, globaliseerumine, linnastumine ja meedia mõju ühiskonnas).

Tekstide keskmine pikkus on A2-tasemel 46,8 sõnet ehk tekstisõna (standardhälve 12,0), B1-tasemel 110,4 sõnet (standardhälve 21,1), B2-tasemel 166,1 sõnet (standardhälve 29,2) ja C1-tasemel 259,8 sõnet (standardhälve 44,25). Autorite vanus jääb vahemikku 13–70, keskmine vanus on 31,8 (standardhälve 13,0) ja mediaanvanus 30. Neist 71,9% on nais- ja 27,1% meessoost, 1% sugu on teadmata. Suurim osa eksaminandidest on kõrgharidusega (41%), ligi veerand (23,3%) alg- või põhiharidusega (enamjaolt kooliõpilased). Sarnasel määral on esindatud keskhariduse (17,3%) ning keskeri- või kutseharidusega inimesed (16,9%). Teave puudub 1,5% eksaminandide haridustaseme kohta.

Eksamisooritajaid on 32 kohalikust omavalitsusest ja kümnest maakonnast. Valdav osa on pärit Harjumaalt (48,8%) ja Ida-Virumaalt (38,3%), enamasti Tallinnast (42,5%), Narvast (20,2%) ja Kohtla-Järvelt (8,1%). Eksaminandide seas on 26 riigi kodakondsusega inimesi. Enim on Eesti (53,1%), Venemaa (20,2%), Ukraina (4,2%), Soome (3,1%), Läti (2,1%) ja määratlemata kodakondsusega ehk välismaalase passiga (10%) eksaminande.

Uurimismaterjal ei võimalda käsitleda õppija emakeele mõju eesti keele kasutusele, sest eksamile tulnud inimestelt ei ole kogutud infot nende emakeele kohta. Kuna Eestis sooritavad tasemeeksameid domineerivalt vene emakeelega inimesed, kes on ka eesti keele kui riigikeele õppe põhisihtrühm, siis kajastab korpus suuresti neile omaseid keelelisi eelistusi. Kõige mitmekülgsema rahvus- ja keeletaustaga on A2-taseme kirjutiste autorid, kelle hulgas on 19 riigi kodakondseid.

Ehkki eksamil osalenud inimeste demograafiline taust on keeleoskustasemeti erinev, on eksamitekstidest koostatud tasakaalustatud valim, mille maht võimaldab teha statistilisi üldistusi ja sisulisi järeldusi eri tasemeile iseloomuliku keelekasutuse kohta.

2.2. Andmeanalüüs

2.2.1. Tekstide automaatanalüüsi põhjal leitud tunnused

Tekstide automaatanalüüsiks on kasutatud Stanfordi ülikoolis arendatud keeletöötlustarkvara Stanza, mis tugineb mahukate keeleandmetega treenitud tehisnärvivõrkudele ja võimaldab analüüsida tekste enam kui 60 keeles, mh määrata sõnade algvormid ehk lemmad, sõnaliigid ja grammatilised vormid. Tekstidele lisatud lingvistiline märgendus on käsitsi kontrollitud ja parandatud, lähtudes Stanza analüüsimudelite treenimisel rakendatud Eesti universaalsete sõltuvuste puudepanga (EstUD) märgenduspõhimõtetest[1]. Saadud andmeid on töödeldud Pythoni andmeanalüüsi teegiga Pandas, et koostada statistiline andmestik eri tasemete tekstide kõrvutamiseks.

Leksikaalsed tunnused (kokku 16) kirjeldavad sõnavara keerukust/rikkust (ingl *lexical complexity*/*richness*), mille põhiaspektidena on välja toodud sõnavara mitmekesisus (*lexical diversity*), ulatus ehk teisisõnu harvaesinevate sõnade esinemus (*lexical sophistication*) ja leksikaalne tihedus ehk sisusõnade osakaal tekstis (*lexical density*) (vt Lu 2012). Lisaks on Jaan Mikk (1979) ja mitmed teised seostanud teksti keerukusega sõnavara abstraktsust (vt Solovyev jt 2020).

Klassikalised sõnavara mitmekesisuse mõõdikud on unikaalsete sõnade arv ja osakaal sõnede suhtes (ingl *type-token ratio*). Granger ja Wynne (1999), samuti Treffers-Daller jt (2018) on soovitanud õppijakeele sõnavara vaheldusrikkust analüüsides käsitleda unikaalsete sõnadena erineva algvormiga sõnu, mitte kõiki sõnavorme. Ka siinses uurimuses vaadeldakse sõnavara mitmekesisust lemmade arvu ja lemmade-sõnede suhtarvu (LSS) alusel. Vähendamaks teksti pikkuse mõju (mida pikem

[1] Stanza väljundis oli korrektselt piiritletud üle 96% lausetest, lemmatiseerimise täpsus oli sõltuvalt tekstide tasemest umbes 92–97% ja sõnaliigi tuvastuse täpsus umbes 97–99% (kõrgematel tasemetel on täpsus suurem). Vormivead kaasnevad reeglina lemmatiseerimis- ja/või sõnaliigi määramise veaga, eraldi tuli neid parandada umbes 0,5–1% sõnede märgenduses. Lähtuti grammatilisest vormist, mida õppija kasutas, mitte sellest, mis vorm olnuks korrektne.

tekst, seda väiksem on erinevate sõnade osakaal) on samast suhtarvust tuletatud rida korrigeeritud indekseid. Siin kasutatakse järgmisi:

- juuritud lemmade-sõnede suhtarv (JLSS; *ingl root type-token ratio*)

$$KLSS = \frac{lemmade\ arv}{\sqrt{2 \times sõnede\ arv}}$$

- korrigeeritud lemmade-sõnede suhtarv (KLSS; ingl *corrected type-token ratio*)

$$KLSS = \frac{lemmade\ arv}{\sqrt{2 \times sõnede\ arv}}$$

- Maasi indeks a^2

$$a^2 = \frac{\log sõnede\ arv - \log lemmade\ arv}{\log sõnede\ arv^2}$$

- Uberi indeks U

$$U = \frac{\log sõnede\ arv^2}{\log sõnede\ arv - \log lemmade\ arv}$$

Eraldi vaadeldakse tegusõnakasutuse mitmekesisust indeksite alusel, mis Vajjala ja Lõo (2014: 122–123) andmetel keeleoskustasemeid hästi eristavad:

- verbivarieeruvuse ruut (RVV; *ingl squared verb variation*)

$$RVV = \frac{tegusõna\ lemmade\ arv^2}{tegusõna\ sõnede\ arv}$$

- korrigeeritud verbivarieeruvus (KVV; ingl *corrected verb variation*)

$$KVV = \frac{tegusõna\ lemmade\ arv}{\sqrt{2 \times tegusõna\ sõnede\ arv}}$$

Kõiki loetletud varieeruvusindekseid peale Maasi indeksi on kasutatud ka eestikeelsete, mh eri keeleoskustaseme tekstide kõrvutamiseks (Pajupuu jt 2009; Alp jt 2013; Kerge jt 2014a; Vajjala & Lõo 2014).

McCarthy ja Jarvise (2007, 2010) katsete põhjal on Maasi indeks logaritmilistest sõnavara mitmekesisuse indeksitest kõige vähem tundlik teksti pikkuse suhtes ning korreleerub tugevalt keerukamate varieeruvusmõõdikutega, mille arvutamiseks jagatakse tekstid mitmeks juhuslikuks või järjestikuseks alamvalimiks (MTLD, vocd-D ja HD-D). Viimased on siinsest analüüsist kõrvale jäetud.

Sõnavara ulatust mõõtvates uurimustes on harvaesineva sõnavarana määratletud sõnu, mis ei kuulu vastava keele 1000–7000 kasutatuima hulka (Lu 2012: 192). Eestis on tekstide sõnavara ulatust hinnates baassõnavaraks loetud Kaalepi ja Muischneki (2002) sagedussõnastiku 3000–4000 sagedamat sõna (nt Pajupuu jt 2009; Kerge jt 2014a). Siinses uurimuses on eri tasemete kirjutistes võrreldud nii 1000, 2000, 3000, 4000 kui ka 5000 sagedama sõna hulka mittekuuluva sõnavara esinemust sõnedes. Selleks on kasutatud päringut Tartu ülikooli andmebaasi[2], mis sisaldab Tasakaalus korpuse sõnasagedusandmeid (Kirt 2013).

Sama andmebaas hõlmab ka eesti keele nimisõnade abstraktsuse hinnanguid kolmeastmelisel skaalal, mille toel on leitud nimisõnade keskmine abstraktsus analüüsitavates tekstides. Hinnangud lähtuvad järgmisest skaalast:

1 – nimisõnad, mis tähistavad meeltega vahetult tajutavaid esemeid ja olendeid;
2 – nimisõnad, mis tähistavad meeltega vahetult tajutavaid nähtusi ja protsesse;
3 – nimisõnad, mis tähistavad meeltega vahetult mittetajutavaid objekte. (Mikk jt 2003: 3)

Leksikaalse tiheduse arvutamisel on sisu- ja funktsioonisõnu eristatud eesti keele stoppsõnade loendi alusel, mis hõlmab side-, ase- ja kaassõnade kõrval ase- ja abimäärsõnu ning abi- ja modaaltegusõnu (Uiboaed 2018). Tinglikult vaadeldakse siinses uurimuses leksikaalsete tunnustena ka nominaalsust, mida saab mõõta nimi- ja tegusõnade suhtarvuna S : V (vt Puksand & Kerge 2012; Pilán 2018), ning formaalsust

[2] Andmebaasile tugineva veebirakenduse on arendanud Marten Siiber (2018). http://prog.keeleressursid.ee/abstraktsus/ (26.10.2021).

ehk ühemõttelisust, mida väljendav F-indeks (Heylighen & Dewaele 2002) on eesti keelele kohandatud (Kerge jt 2007):

F = (*nimisõnad* (%) + *omadussõnad* (%) + *kaassõnad* (%) – *asesõnad* (%) – *tegusõnad* (%) – *määrsõnad* (%) – *hüüdsõnad* (%) + 100) ÷ 2

Üheselt mõistetavale, vähe kontekstile tuginevale formaalsele esituslaadile vastandub kontekstuaalsus ehk mitmemõttelisus, mis omasem mitteametlikule ja vahetule suhtlusele. Leksikaalne tihedus, nominaalsus ja formaalsus seonduvad teksti infotihedusega ja rajanevad sõnaliikide sagedussuhetel, iseloomustades ühtlasi ka keeleõppija sõnakasutusega seotud grammatikat.

Morfoloogiliste tunnuste (kokku 53) valikul on lähtutud kategooriatest, mida Stanza abil lisatud märgendus võimaldab automaatanalüüsis arvesse võtta. Tunnused jagunevad kolme rühma.

1) Sõnaliigitunnused (19) hõlmavad nimi-, omadus-, ase-, arv-, tegu-, määr-, side-, kaas- ja hüüdsõnade osakaalu tekstis, lisaks nende mõningate alamliikide sagedust: rinnastavad ja alistavad sidesõnad, ees- ja tagasõnad, pärisnimed ning isikulised, enesekohased, näitavad, umbmäärased ja küsivad-siduvad asesõnad.
2) Käändsõnatunnused (18) iseloomustavad summaarselt nimi-, omadus-, ase- ja arvsõnade kasutust. Arvutatakse 14 võimaliku käändevormi (sisseütleva pikka ja lühikest vormi vaadeldakse vähese esinemuse tõttu koos) ning ainsuse- ja mitmusevormide osakaal teksti sõnede suhtes, loendatakse tekstis esinevate käändevormide arvu.
3) Tegusõnatunnused (17) kajastavad pöördeliste vormide, kolme kõneviisi (kindel, tingiv ja käskiv), 1., 2. ja 3. pöörde vormide, ainsuse- ja mitmusevormide, oleviku- ja lihtminevikuvormide, käändeliste vormide (ka eraldiseisvalt tegevusnimede, *des*-vormi, mineviku kesksõnade), eitusvormi kuuluvate sõnade (partikkel ja tegusõnavorm) ning umbisikulise tegumoe vormide osakaalu tekstis.

Siinsel analüüsietapil ei vaadelda eraldi sõnade vormikasutust käändsõnaliikide lõikes ega ka käänd- ja tegusõnatunnuste koosesinemise sagedust. Grammatiliste vormide osakaalud on sarnaselt Vajjala ja Lõoga (2013, 2014) arvutatud kõigi sõnede suhtes, ehkki vormisagedust võib mõõta ka käänd- ja tegusõnade arvu suhtes (nt Hancke 2013; Pilan 2018; Szügyi jt 2019).

2.2.2. Statistilised meetodid

Et leida keeleoskustasemeid oluliselt eristavad lingvistilised tunnused ja vaadelda nende tunnuste koosesinemist, on rakendatud Welchi ANOVA olulisustesti, korrelatsioonanalüüsi ja mitmemõõtmelist skaleerimist, kasutades tarkvarapaketti SPSS Statistics (versioon 25.0).

Tasemete erinevuste hindamiseks on traditsioonilise dispersioonanalüüsi asemel valitud Welchi ANOVA ehk parandatud F-statistik, mis ei eelda tunnuste sarnast hajuvust (varieerumist) võrreldavates rühmades – siinses andmestikus ei ole see eeldus enamasti täidetud, st et tunnuste väärtused kõiguvad tasemeti erineval määral. Üle poole tunnustest (37) vastab kõigil keeleoskustasemetel ligilähedaselt normaaljaotusele, ülejäänud tunnuste jaotus kaldub ühel või mitmel tasemel väiksemate väärtuste poole – enamjaolt on need sõnaliigid ja grammatilised vormid, mis esinevad madalamatel tasemetel väga harva. Welchi ANOVA on aga normaaljaotuse eelduse mittetäidetuse suhtes vähetundlik (Delacre jt 2019).

Erinevuse statistilise olulisuse määramisel lähtutakse olulisusnivoost 5% ($\alpha = 0{,}05$). Kuna nelja taseme tekste kõrvutatakse 69 tunnuse alusel ja olulisustesti korduv rakendamine sama grupeeriva tunnuse puhul suurendab juhuslike eksimuste tõenäosust, siis on iga eraldiseisva testi olulisusnivood korrigeeritud Bonferroni meetodil: $\alpha \div n$, kus n on testitavate tunnuste arv (vt Armstrong 2014). Korrigeeritud olulisusnivoo on ümardatud kolme komakohani: $0{,}05 \div 69 = 0{,}0007 \sim 0{,}001$. Niisiis on tasemete erinevus loetud statistiliselt oluliseks, kui olulisustõenäosus $p \leq 0{,}001$.

Sel viisil oluliseks osutunud tunnused on võetud aluseks keeleoskustasemete paarikaupa võrdlemisel, milleks on kasutatud Gamesi-Howelli järeltesti. See meetod arvestab korduva testimisega (nt nelja võrreldava rühma puhul kuus paarikaupa võrdlust) ja võimaldab arvutada korrigeeritud p-väärtuse valitud olulisusnivool (siin $\alpha = 0{,}05$). Kahe taseme erinevus loetakse oluliseks, kui $p \leq 0{,}05$. Siinses artiklis tuuakse esile järjestikuste keeleoskustasemete erinevused, mida võib taseme automaatsel hindamisel pidada esmatähtsaks.

Welchi F-statistikust on lähtutud ka sama taseme eksamite erinevuste tuvastamisel, et leida kirjutamisülesandest sõltuvad tunnused. Nii on kõiki tunnuseid kõrvutatud iga taseme eksamikordade lõikes, võttes kõigi nelja taseme puhul aluseks taas Bonferroni meetodil korrigeeritud olulisusnivoo ~ 0,001. Artiklis kirjeldatakse suuremaid tasemesiseseid erinevusi. Järeltesti ei ole rakendatud.

Tasemetevahelisi erinevusi ilmestavad tunnuste usaldusvahemikud ehk üldkogumi keskväärtuste hinnangud iga taseme jaoks (siin tõenäosusega 95%), mis põhinevad valimi kohta arvutatud aritmeetilistel keskmistel ja standardhälvetel. Juhul, kui usaldusvahemikud kattuvad, ei ole erinevus tasemete vahel oluline.

Tunnuste lähedus-kaugus ning seos õppija kirjaliku keeleoskuse tasemega (madalam/kõrgem) esitatakse visuaalselt mitmemõõtmelise skaleerimise meetodil ALSCAL-algoritmi kasutades. Samade lingvistiliste tunnuste vahelisi seoseid tõlgendatakse Pearsoni lineaarse korrelatsioonikordaja (r) väärtuste alusel, lähtudes Rowntree (1981) pakutud jaotusest: < 0,2 olematu; 0,2–0,4 nõrk; 0,4–0,7 keskmine; 0,7–0,9 tugev; > 0,9 väga tugev. Välja tuuakse vähemalt keskmise tugevusega seosed.

3. Keeleoskustasemeid eristavad tunnused

Statistiliselt oluliste erinevuste ilmnemine keeleoskustasemete vahel võimaldab lingvistilised tunnused liigitada järgmiselt:

1) keeleoskustasemeid läbivalt eristavad tunnused, mis piiritlevad tasemeid A2–B1, B1–B2 ja B2–C1;

2) kahte järjestikuste tasemete paari eristavad tunnused, mis piiritlevad tasemeid a) A2–B1 ja B1–B2, b) A2–B1 ja B2–C1 või c) B1–B2 ja B2–C1;
3) ühte järjestikuste tasemete paari eristavad tunnused, mis piiritlevad tasemeid a) A2–B1, b) B1–B2 või c) B2–C1;
4) tunnused, mis järjestikuseid tasemeid ei erista (sh tunnused, mille väärtus on kõigil tasemetel sarnane).

Toodud liigituse piires võib tunnuste avaldumise dünaamika olla lineaarne (samasuunaliselt kasvav või kahanev) või mittelineaarne ehk erisuunaline. Järgnev kirjeldus keskendub vähemalt kahte tasemepaari eristavatele tunnustele. Welchi ANOVA tulemused on esitatud lisas 1, kus on välja toodud ja vastavalt tähistatud ka ühte tasemepaari ja mittejärjestikuseid tasemeid eristavad tunnused.

3.1. Leksikaalsed tunnused

Leksikaalsete tunnuste hulgas leidub nii keeleoskustasemeid läbivalt eristavaid tunnuseid (10) kui ka neid, mis eristavad kahte või ühte tasemepaari (vastavalt kaks ja viis tunnust, vt tabel 1 ja lisa 1). Tabelis 1 on tasemete kaupa esitatud olulisemate eristavate tunnuste aritmeetiline keskmine ja standardhälve valimis ning keskväärtuse usaldusvahemik. Gamesi-Howelli testi tulemusel saadud korrigeeritud p-väärtus näitab erinevuste statistilist olulisust järjestikuste tasemete võrdluses.

Tabelist 1 ilmneb viis tendentsi.

Kõiki järjestikuseid tasemeid eristavad peamiselt sõnavara mitmekesisust iseloomustavad tunnused. Ootuspäraselt suureneb tase-tasemelt lemmade ehk erineva algvormiga sõnade hulk, mis kasvab koos teksti pikkusega. Samuti suurenevad lemmade-sõnede suhtarvust (LSS) tuletatud klassikalised indeksid JLSS ja KLSS (mõlemad väikese hajuvusega) ning tegusõnade varieeruvust kirjeldavad indeksid RVV ja KVV, millest viimasel on suhteliselt väike hajuvus. Need tunnused on aga tugevalt seotud lemmade arvuga tekstis.

Seevastu ei kajasta logaritmilised Maasi ja Uberi indeks sõnavara järkjärgulist avardumist: nende alusel varieerub sõnavara B1-tasemel vähem kui A2-tasemel ning B2-tasemel sarnaneb leksikaalne mitmekesisus A2-tasemega, suurenedes alles C1-tasemel. Niisuguse mittelineaarse muutuse põhjust võib otsida selles, et B1-tasemel on tekstid keskmiselt 2,4 korda pikemad, kuid lemmade arv suureneb 1,9 korda. Järelikult sõnavara küll täieneb, ent tekstide pikkus suureneb kiiremini kui neis esinevate unikaalsete sõnade hulk. B2-tasemel suureneb sõnede ja lemmade arv ühtviisi 1,5 korda ning C1-tasemel 1,6 korda, mis tähendab, et tekstid pikenevad pigem unikaalse sõnavara arvelt. Samas on mõlemad indeksid igal tasemel suurema hajuvusega kui JLSS ja KLSS.

Tabel 1. *Keeleoskustasemeid eristavad leksikaalsed tunnused ja tasemetevahelise erinevuse statistiline olulisus (olulisusnivoo 0,05)*

Tunnus	Tase	Keskväärtus	Standardhälve	Usaldusvahemik tõenäosusega 95%	p-väärtus
Eristavad keeleoskustasemeid läbivalt					
Lemmade arv	A2	33,4	6,8	32,2…34,6	< 0,001
	B1	63,0	11,9	60,9…65,2	< 0,001
	B2	94,4	15,4	91,6…97,1	< 0,001
	C1	147,6	23,1	143,5…151,8	
JLSS	A2	4,7	0,5	4,6…4,8	< 0,001
	B1	5,8	0,7	5,7…5,9	< 0,001
	B2	7,2	0,7	7,0…7,3	< 0,001
	C1	9,1	1,0	8,9…9,3	
KLSS	A2	3,3	0,4	3,2…3,4	< 0,001
	B1	4,1	0,5	4,0…4,2	< 0,001
	B2	5,1	0,5	5,0…5,2	< 0,001
	C1	6,4	0,7	6,3…6,6	
Maasi indeks	A2	0,024	0,007	0,022…0,025	0,002
	B1	0,027	0,005	0,026…0,028	< 0,001
	B2	0,022	0,004	0,022…0,023	< 0,001
	C1	0,018	0,004	0,018…0,019	

Uberi indeks	A2 B1 B2 C1	47,9 39,1 46,2 56,7	22,1 8,4 7,9 12,0	43,9…51,9 37,5…40,6 44,8…47,6 54,6…58,9	< 0,001 < 0,001 < 0,001
RVV	A2 B1 B2 C1	5,9 8,0 12,7 17,7	2,1 3,1 3,9 5,3	5,6…6,3 7,4…8,5 12,0…13,4 16,7…18,6	< 0,001 < 0,001 < 0,001
KVV	A2 B1 B2 C1	1,7 2,0 2,5 2,9	0,3 0,4 0,4 0,4	1,6…1,8 1,9…2,0 2,4…2,6 2,9…3,0	< 0,001 < 0,001 < 0,001
Leksikaalne tihedus	A2 B1 B2 C1	53,9 50,2 47,9 57,0	7,6 6,3 6,1 5,5	52,6…55,3 49,1…51,4 46,8…49,0 56,0…58,0	< 0,001 0,017 < 0,001
Harvad lemmad (%, mitte 5000 sagedama seas)	A2 B1 B2 C1	17,8 12,3 16,3 20,1	6,7 4,3 5,2 4,3	16,5…19,0 11,5…13,0 15,4…17,2 19,3…20,9	< 0,001 < 0,001 < 0,001
Harvad lemmad (%, mitte 4000 sagedama seas)	A2 B1 B2 C1	20,9 17,2 20,0 23,5	7,1 5,8 5,4 4,5	19,6…22,1 16,1…18,2 19,1…21,0 22,7…24,3	< 0,001 0,001 < 0,001
Eristavad tasemeid B1–B2 ja B2–C1					
Nomi-naalsus (S : V)	A2 B1 B2 C1	1,3 1,3 1,1 1,5	0,4 0,6 0,4 0,4	1,2…1,4 1,2…1,4 1,0…1,2 1,4…1,6	0,986 0,017 < 0,001
Nimisõnade keskmine abstraktsus-hinnang	A2 B1 B2 C1	1,6 1,6 1,9 2,3	0,2 0,3 0,3 0,2	1,5…1,6 1,6…1,7 1,8…1,9 2,2…2,3	0,357 < 0,001 < 0,001

Korrigeerimata LSS, mis on teksti pikkuse mõjule kõige tundlikum, väheneb B1-tasemel ja püsib järgnevatel tasemetel sarnane, kuna tekstide pikkus ja neis esindatud lemmade arv kasvavad võrdsel määral.

Sõnavara ulatus suureneb eelkõige vilunud keelekasutaja kirjutistes. 1000, 2000 ja 3000 sagedama sõna hulka mittekuuluva sõnavara kasutus eristab üksnes C1-taset. 4000 ja 5000 sagedama sõna hulka mittekuuluva sõnavara esinemus eristab tasemeid läbivalt, kuid mitte lineaarselt. Sellise sõnavara osakaal väheneb B1-tasemel ja suureneb C1-tasemeni, kusjuures A2- ja B2-taseme võrdluses statistiliselt olulisel määral ei erine. Nagu eespool välja toodud, suureneb B1-tasemel tekstide pikkus kiiremini kui sõnavara mitmekesisus, mis võib mõjutada ka sõnavara ulatust. Lisaks võib harvaesineva sõnavara väiksem osakaal B1-tasemel tuleneda kirjutamisülesandest. A2-taseme lühikirjades räägitakse reisidest-väljasõitudest ja automarkidest ning viidatakse kirja adressaadile – seetõttu kasutatakse rohkem pärisnimesid, mis lähevad arvesse harvade sõnadena (keskmiselt 6,5%, B1-tasemel 3,7% sõnedest).

Teksti leksikaalne tihedus väheneb B2-tasemeni, suurenedes märkimisväärselt C1-tasemel. Sarnaselt muutub tekstide **nominaalsus** ja **formaalsus**, ehkki nimi- ja tegusõnade suhtarv eristab statistiliselt olulisel määral B1–C1-taset ning F-indeks vaid B2- ja C1-taset. See suundumus on seotud peamiselt nimisõnade kasutusega, mis harveneb B2-tasemeni (25,6%) ja suureneb järsult C1-tasemel (33,0%). Sisusõnade osakaalu võib vähendada ka tekstisidususvahendite tarvitamine. A2–B2-tasemel laieneb sidesõnade kasutus ning B1-tasemel määrsõnade kasutus – enim tarvitatakse väga üldise tähendusega määrsõnu, mis sisalduvad stoppsõnade loendis (*väga*, *ka*, *palju*, *praegu*, *veel*, *seal*, *siis, koos*, *nii*, *juba*). C1-taseme tekstide suurem leksikaalne tihedus ja formaalsus tulenevad asesõnade osakaalu järsust langusest (B2 18,9%; C1 12,6%).

Kuna leksikaalset tihedust on seni arvutatud sõnaliikide sageduse alusel, mitte stoppsõnade loendi abil, siis ei saa siinseid andmeid võrrelda varasemate uurimistulemustega (nt Kerge jt 2014a). A2- ja B1-taseme eksamikirjutiste nominaalsus on Kerge jt (2014b: 60) andmetele

tuginedes sarnane 9. ja 11. klassi õpilaste kirjutistega (1,3), B2-taseme kirjutised on lähedasemad suulisele dialoogile (1,1) ja C1-taseme kirjutised haritlaste esseedele (1,5). Siin avaldub isiklike ja vabamas vormis ametikirjade suurem lähedus suulisele ja dialoogilisele keelekasutusele võrreldes teiste kirjaliku teksti liikidega (vt Kerge 2010).

Kirjutamisülesandega seostub ka tekstide formaalsus (keskmine F-indeks ja standardhälve: A2-tasemel 42,9 ja 9,4; B1-tasemel 40,9 ja 6,2; B2-tasemel 38,9 ja 6,6; C1-tasemel 48,2 ja 5,9). Võttes aluseks Puksandi ja Kerge (2012: 182) esitatud suulise ja kirjaliku teksti žanrite F-indeksi mediaanväärtused ning eri keeleoskustasemete F-indeksi mediaanid (erinevad väga vähe aritmeetilistest keskmistest), saab väita, et C1-taseme arutluste formaalsus (mediaan 48,5) sarnaneb kirjalike žanritega (mediaan 49,6); A2- ja B1-taseme teadete-jutustuste formaalsus (mediaan vastavalt 42,7 ja 41,5) on lähedane suuliste ja kirjalike žanrite üldisele keskmisele (mediaan 43,2); B2-taseme kirjade formaalsus (mediaan 38,9) on sellest väiksem, kuid suurem võrreldes suuliste žanritega (mediaan 34,8). Žanrivõrdlus on näidanud, et arutlevad tekstid (esseed) on formaalsemad kui nt ilukirjandus, erakirjad ja ettevõttesisene kirjavahetus (Pajupuu & Kerge 2010: 386).

Nimisõnade abstraktsus eristab B1–C1-taset. Enamikus A2- ja B1-taseme tekstides vastab suurem osa nimisõnadest abstraktsuse astmele 1 (nt *sõber*, *kodu*, *kohvik*), kuid peaaegu kõigis kirjutistes leidub ka selliseid, mille abstraktsuse aste on 2 (nt *reis*, *õhtu*, *muusika*). Nimisõnu, mis tähistavad meeltega vahetult tajumatuid objekte (aste 3), tuleb esile vähe: korduvad eelkõige sõnad *aasta* ja *aeg*, mis sagedad B2- ja C1-tasemel ning eesti kirjakeeleski. A2-tasemel kasutatakse ka abstraktseid sõnu *abi*, *hind*, *aadress* ning kuude ja nädalapäevade nimetusi (nt *juuli*, *esmaspäev*), B1-tasemel sõnu *info*, *teema*, *sport*, *elu*, *kuu*, *nädal* ja *maailm*. Kõige ühtlasemalt jagunevad nimisõnad kolme abstraktsuse astme vahel B2-tasemel. Sagedate abstraktsete nimisõnadena on kasutusel *probleem*, *vastus*, *mõte*, *olukord*, *võimalus*, *lahendus*, *nõu, lugupidamine*. C1-tasemel on kõrgeima abstraktsuse astmega nimisõnad harilikult ülekaalus, nt *riik*, *rahvus*, *kultuur*, *ühiskond*, *põhjus*, *arvamus*, *kasu*, *huvi*. See on

oodatav tulemus, sest kirjutamisülesanne eeldab arutlemist abstraktsetel teemadel nagu sallivus, tervislik eluviis, kodanikualgatus.

Eksamite lõikes erinevad kõige vähem sõnavara mitmekesisuse tunnused. Üksnes RVV ja KVV erinevad olulisel määral A2-taseme eksamikordade võrdluses. A2–B2-tasemel sõltub eksamiülesandest tekstide leksikaalne tihedus, nominaalsus ja formaalsus, samuti sõnavara ulatus. B1–C1-taseme eksamite lõikes erineb nimisõnade keskmine abstraktsus.

Sõnavara valikut mõjutab nii kirjutiste teema kui ka tekstiliik. Nt B1-tasemel on suurima leksikaalse tihedusega, kõige formaalsemad, nominaalsemad ja abstraktsema sõnavaraga meediakasutust käsitlevad eksamitekstid, mis loetlevad autorile huvipakkuvaid valdkondi ja infokanaleid. Kõige verbirikkamad ja vähem abstraktsed on tekstid lemmikloomade pidamisest, kus fookus langeb lemmikuga seotud tegevustele ja soovidele, nt *mängima*, *jalutama*, *jooksma*, *õpetama*, *hoolitsema*, *söötma*, *koristama*, *tahtma*, *unistama*, *võtma*, *valima*. Harvaesineva sõnavara osakaal on suurim meedia- ja kontserditeemalistes tekstides, kus tarvitatakse rohkem spetsiifilist sõnavara (*portaal*, *koduleht*, *popmuusika*, *kammerkoor*).

B2-tasemel avaldub erinevus ametlike ja mitteametlike kirjade keelekasutuses. Kolleegile või kohalikule omavalitsusele suunatud ametikirjad on leksikaalselt tihedamad, nominaalsemad ja formaalsemad kui kirjad, kus küsitakse tuttavalt nõu tööalase või isikliku probleemi lahendamiseks. Ka harvaesineva sõnavara osakaal on suurem ametikirjades.

Järeldusi. Sõnavara mitmekesisus suureneb kõigil keeleoskustasemetel, ent unikaalseid sõnu lisandub rohkem B2- ja C1-tasemel. Nagu ka varasemate uurimuste alusel (Alp jt 2013; Vajjala & Lõo 2014) kindlaks tehtud, eristavad eesti keele õppijate A2–C1-taseme kirjutiste sõnavara mitmekesisust läbivalt indeksid JLSS (nimetatud ka Guiraud' indeksiks) ja KLSS. Nende puhul tuleb aga silmas pidada tugevat seost teksti pikkusega.

Harvaesineva sõnavara osakaal suureneb peamiselt C1-tasemel. Siiski kuulub ka vilunud keeleõppija tekstides umbes 80% sõnadest eesti

keele 5000 sagedama hulka. Samuti eristuvad C1-taseme tekstid selgelt leksikaalse tiheduse, nominaalsuse ja formaalsuse poolest: arvamustekst eeldab viitetihedamat ja ühemõttelisemat esituslaadi kui isiklikud kirjad, vabamas vormis ametikirjad ja jutustavad tekstid.

Automaatse tasemehindamise jaoks on relevantsemad samasuunaliselt kasvavad sõnavara mitmekesisuse mõõdikud ja nimisõnade abstraktsus.

3.2. Morfoloogilised tunnused

Morfoloogiliste tunnuste hulgas (kokku 53) tõuseb esile 12 keeleoskustasemeid läbivalt eristavat tunnust ja 18 kahte järjestikuste tasemete paari piiritlevat tunnust (vt tabel 2, ülejäänud tunnuste kohta vt lisa 1).

Tabel 2. *Keeleoskustasemeid eristavad morfoloogilised tunnused ja tasemetevahelise erinevuse statistiline olulisus (olulisusnivoo 0,05)*

Tunnus	Tase	Keskväärtus	Standardhälve	Usaldusvahemik tõenäosusega 95%	p-väärtus
Eristavad keeleoskustasemeid läbivalt					
Alistavad sidesõnad (%)	A2 B1 B2 C1	1,3 2,9 4,1 3,5	1,9 2,1 1,4 1,3	1,0…1,7 2,5…3,3 3,8…4,3 3,2…3,7	< 0,001 < 0,001 0,003
Hüüdsõnad (%)	A2 B1 B2 C1	1,3 0,1 0,5 0,0	1,6 0,3 0,5 0,1	1,0…1,6 0,1…0,2 0,4…0,6 0,0…0,05	< 0,001 < 0,001 < 0,001
Pärisnimed (%)	A2 B1 B2 C1	6,5 3,7 2,2 1,1	4,3 3,2 1,7 1,3	5,7…7,2 3,1…4,3 1,9…2,5 0,8…1,3	< 0,001 < 0,001 < 0,001
Käändevormide arv	A2 B1 B2 C1	6,3 8,2 9,2 10,6	1,3 1,3 1,4 1,2	6,1…6,6 8,0…8,4 8,9…9,4 10,4…10,8	< 0,001 < 0,001 < 0,001

Tunnus	Tase	Keskväärtus	Standardhälve	Usaldusvahemik tõenäosusega 95%	p-väärtus
Käändsõnad seestütlevas käändes (%)	A2	0,4	0,9	0,2…0,5	< 0,001
	B1	1,7	2,1	1,3…2,1	0,005
	B2	1,0	1,1	0,8…1,2	< 0,001
	C1	2,2	1,2	2,0…2,4	
Käändsõnad ainsuses (%)	A2	52,2	7,9	50,8…53,7	< 0,001
	B1	45,2	6,8	43,9…46,4	< 0,001
	B2	38,4	7,1	37,1…39,6	0,001
	C1	35,3	5,1	34,3…36,2	
Käändsõnad mitmuses (%)	A2	4,6	4,7	3,8…5,5	< 0,001
	B1	9,2	5,1	8,3…10,1	< 0,001
	B2	12,4	6,1	11,3…13,5	< 0,001
	C1	18,3	4,0	17,6…19,1	
Tegusõnad käskivas kõneviisis (%)	A2	1,0	1,5	0,7…1,3	< 0,001
	B1	0,2	0,5	0,1…0,3	0,001
	B2	0,5	0,7	0,4…0,6	< 0,001
	C1	0,1	0,4	0,0…0,2	
Tegusõnad 1. pöördes (%)	A2	8,9	5,8	7,8…9,9	0,045
	B1	7,3	2,7	6,8…7,8	< 0,001
	B2	5,1	1,9	4,7…5,4	< 0,001
	C1	1,8	1,4	1,6…2,1	
Tegusõnad 2. pöördes (%)	A2	2,3	2,7	1,8…2,7	< 0,001
	B1	0,5	1,0	0,3…0,6	< 0,001
	B2	1,6	1,5	1,3…1,9	< 0,001
	C1	0,2	0,5	0,1…0,3	
Tegusõnad 3. pöördes (%)	A2	7,8	4,1	7,1…8,6	< 0,001
	B1	9,8	3,1	9,2…10,3	< 0,001
	B2	7,5	2,6	7,0…8,0	< 0,001
	C1	10,1	2,2	9,7…10,5	
Tegusõna eitusvormid (%)	A2	0,9	1,8	0,6…1,2	0,010
	B1	1,6	1,8	1,3…1,9	< 0,001
	B2	3,7	2,5	3,3…4,2	< 0,001
	C1	2,2	1,6	1,9…2,5	
Eristavad tasemeid A2–B1 ja B1–B2					
Küsivad-siduvad asesõnad (%)	A2	0,1	0,6	0…0,2	< 0,001
	B1	0,6	1,1	0,4…0,8	< 0,001
	B2	1,3	1,1	1,1…1,5	0,410
	C1	1,5	0,8	1,3…1,6	

Tunnus	Tase	Keskväärtus	Standardhälve	Usaldusvahemik tõenäosusega 95%	p-väärtus
Sidesõnad (%)	A2	5,1	3,2	4,5…5,7	< 0,001
	B1	8,0	2,9	7,5…8,5	< 0,001
	B2	9,5	2,1	9,1…9,9	0,079
	C1	8,9	1,7	8,6…9,2	
Eristab tasemeid A2–B1 ja B2–C1					
Arvsõnad (%)	A2	3,6	4,5	2,8…4,5	< 0,001
	B1	1,3	1,4	1,0…1,6	0,835
	B2	1,4	1,2	1,2…1,7	< 0,001
	C1	0,7	0,7	0,6…0,8	
Eristavad tasemeid B1–B2 ja B2–C1					
Nimisõnad (%)	A2	29,9	7,2	28,6…31,2	0,067
	B1	27,8	6,3	26,6…28,9	0,029
	B2	25,6	5,5	24,6…26,6	< 0,001
	C1	33,0	4,9	32,2…33,9	
Omadussõnad (%)	A2	6,6	3,8	5,9…7,3	0,994
	B1	6,7	2,9	6,2…7,3	0,047
	B2	5,8	2,3	5,4…6,3	< 0,001
	C1	8,2	2,3	7,8…8,6	
Tegusõnad (%)	A2	24,3	4,1	23,5…25,0	0,147
	B1	23,2	3,7	22,5…23,9	0,035
	B2	24,4	3,3	23,8…25,0	< 0,001
	C1	22,5	2,8	22,0…23,0	
Käändsõnad nimetavas käändes (%)	A2	29,3	7,1	28,1…30,6	0,767
	B1	28,5	5,4	27,6…29,5	< 0,001
	B2	21,4	4,2	20,6…22,1	< 0,001
	C1	18,4	3,9	17,7…19,1	
Käändsõnad omastavas käändes (%)	A2	6,7	4,3	6,0…7,5	0,411
	B1	6,0	2,8	5,5…6,5	< 0,001
	B2	9,7	4,4	8,9…10,5	< 0,001
	C1	11,9	3,3	11,3…12,5	
Käändsõnad seesütlevas käändes (%)	A2	3,8	4,4	3,0…4,6	0,351
	B1	3,1	2,0	2,7…3,5	0,001
	B2	2,2	1,6	1,9…2,5	0,001
	C1	3,1	1,7	2,8…3,4	

Tunnus	Tase	Keskväärtus	Standardhälve	Usaldusvahemik tõenäosusega 95%	p-väärtus
Käändsõnad saavas käändes (%)	A2	0,1	0,4	0…0,1	0,054
	B1	0,2	0,4	0,1…0,3	< 0,001
	B2	0,7	1,0	0,5…0,9	< 0,001
	C1	1,9	1,3	1,7…2,2	
Isikulised asesõnad (%)	A2	14,9	6,8	13,6…16,1	0,612
	B1	14,0	3,8	13,3…14,7	< 0,001
	B2	10,7	3,9	10,0…11,4	< 0,001
	C1	3,0	2,0	2,6…3,4	
Näitavad asesõnad (%)	A2	1,9	2,0	1,5…2,3	0,273
	B1	2,3	1,8	2,0…2,7	< 0,001
	B2	3,9	1,7	3,6…4,2	0,012
	C1	4,7	1,9	4,3…5,0	
Tegusõna pöördelised vormid (%)	A2	19,5	4,1	18,8…20,2	0,228
	B1	18,7	2,7	18,2…19,1	< 0,001
	B2	17,0	2,1	16,6…17,4	< 0,001
	C1	14,8	1,8	14,4…15,1	
Tegusõnad kindlas kõneviisis (%)	A2	18,3	4,3	17,5…19,0	0,971
	B1	18,1	2,7	17,6…18,6	< 0,001
	B2	15,5	2,4	15,1…16,0	< 0,001
	C1	13,7	1,9	13,3…14,0	
Tegusõnad ainsuses (%)	A2	14,7	5,2	13,8…15,7	0,703
	B1	14,1	2,9	13,6…14,7	< 0,001
	B2	11,0	2,9	10,5…11,6	< 0,001
	C1	8,2	1,9	7,9…8,6	
Tegusõnad umbisikulises tegumoes (%)	A2	0,2	0,6	0,0…0,3	0,992
	B1	0,2	0,5	0,1…0,3	< 0,001
	B2	0,7	0,8	0,6…0,9	< 0,001
	C1	1,4	1,2	1,2…1,6	
Tegusõna käändelised vormid (%)	A2	4,3	3,7	3,6…5,0	0,429
	B1	3,7	2,7	3,2…4,2	< 0,001
	B2	5,6	1,8	5,3…5,9	< 0,001
	C1	6,7	2,0	6,4…7,1	
da- ja *ma*-tegevusnime vormid (%)	A2	4,1	3,7	3,4…4,8	0,258
	B1	3,3	2,6	2,9…3,8	< 0,001
	B2	4,7	1,7	4,4…5,0	0,005
	C1	5,5	1,9	5,1…5,8	

Kuigi uurimuse eesmärk ei olnud kõrvutada õppijakeelt ja emakeelekõnelejate tekstiloomet, on sõnaliigisageduste ja käändsõna grammatiliste kategooriate kasutuse tõlgendamisel saadud tuge Tasakaalus korpuse statistikast (Sõnaliikide sagedusloend…). Sealne tekstimaterjal on märgendatud täisautomaatselt ja selleks on kasutatud teist tarkvara (Filosofti morfoloogilist analüsaatorit ja statistilist ühestajat t3mesta), ent tegemist on mahuka ja esindusliku tekstikoguga, mis kajastab eesti kirjakeelele omaseid keelekasutustendentse.

3.2.1. Sõnaliikide jaotus

A2–C1-taset eristab läbivalt hüüdsõnade, pärisnimede ja alistavate sidesõnade keskmine osakaal tekstis. Hüüdsõnu kasutatakse enim A2-tasemel. See tuleneb ühelt poolt kirjutamisülesandest (kiri algab tavaliselt hüüdsõnaga *tere*), teisalt teksti pikkusest (nt 50-sõnelise teksti mahust moodustab juba üksainus hüüdsõna 2%). Ka B2-taseme kirjades esineb rohkem hüüdsõnu (*tere* kõrval ka *aitäh* ja *palun*) kui B1-taseme jutustustes ja C1-taseme arutlustes, kus neid kohtab äärmiselt harva.

Pärisnimede osakaal kahaneb tase-tasemelt, sest tekstid pikenevad, kuid nimede absoluutarv muutub vähe. See tuleneb samuti kirjutamisülesandest: C1-taseme arutlused sisaldavad vähem isikunimesid kui eelnevate tasemete kirjad-jutustused.

Alistavate sidesõnade osakaal suureneb B2-tasemeni ja väheneb C1-tasemel. Sarnaselt muutub kõikide sidesõnade osakaal, ehkki erinevus B2- ja C1-taseme vahel pole statistiliselt oluline. Sidesõnade ohtram kasutus viitab keerukamale lausestruktuurile – tarvitatakse rohkem liitlauseid, sh põimlauseid. See, et C1-tasemel alistavate sidesõnade osakaal teksti sõnede suhtes uuesti väheneb, on ilmselt seotud pikemate lausete moodustamisega. Sidesõnade osakaal C1-taseme kirjutistes (keskmiselt 8,9%) on samaväärne Tasakaalus korpuse andmetega (8,2%, alistavate sidesõnade kohta info puudub).

A2–B2-tasemel suureneb küsivate-siduvate asesõnade osakaal. A2-taseme kirjutistes esinevad üksikjuhtudel asesõnad *mis* ja *kes*.

B1-tasemel leidub kõrvallause sidendi või küsisõna funktsioonis tarvitatavaid asesõnu rohkem kui kolmandikus ja B2-tasemel juba enamikus tekstidest, lisanduvad asesõnad *milline* ja *missugune*.

B1–C1-taseme lõikes erineb nimi-, omadus- ja tegusõnade, isikuliste ja näitavate asesõnade kasutus. Nimisõnade osakaal kahaneb B2-tasemeni (A2- ja B1-taset see oluliselt ei piiritle) ning kasvab C1-tasemel hüppeliselt (vahe B2-tasemega 7,4%). Analoogne on omadussõnade kasutuse dünaamika: B2-tasemel nende osakaal mõnevõrra väheneb, olles suurim C1-tasemel. Tegusõnade osakaal muutub vähe: suureneb pisut B2-tasemel ja väheneb C1-tasemel. Kirjeldatud muutused meenutavad siksak-mustrit, mis on tõenäoliselt tingitud eksamikirjutiste tekstiliigist.

Arutlused (C1-tase) eeldavad formaalsemat, st nimi- ja omadussõnarohkemat keelekasutust kui kirjad (B2-tase) ja jutustused (B1-tase). B1- ja B2-taseme erinevus tuleneb suurest kõikumisest sõnaliikide sageduses ja formaalsuses. Kahelt B2-eksamilt pärit mitteametlikud kirjad on oluliselt kontekstitundlikumad (F-indeks 34,7 ja 34,6) kui teiste eksamite ametlikud kirjad (F-indeks 44,7 ja 41,4). C1-taseme tekstides on nimi-, omadus- ja tegusõnade osakaal (vastavalt 33,0%, 8,2% ja 22,5%) väga sarnane Tasakaalus korpuse andmetega (33,7%, 8,2% ja 22,2%). Võib eeldada, et arutlevad publitsistlikku laadi tekstid esindavad nimetatud sõnaliikide sageduse osas eesti kirjakeele keskseid tendentse.

Isikuliste asesõnade osakaal B1–C1-taseme tekstides väheneb, samas kui näitavate asesõnade kasutus laieneb. Isikuliste asesõnade valik sõltub suhtluseesmärgist ja tekstiliigist, mõjutades ka tegusõnavormide kasutust (vt 3.2.3.). Kui kirjutiste teemad muutuvad üldisemaks ja abstraktsemaks ning esineb vähem *mina*-kesksust, siis isikuliste asesõnade sagedus taandub. Näitavaid asesõnu (valdavalt *see*) tarvitatakse viiteseose loomiseks eelneva infoga, mis on omane keerukamale ja sidusamale lausestusele.

Tasemete A2–B1 ja B2–C1 võrdluses väheneb arvsõnade osakaal, mis samuti sõltub eksamikirjutise teemast. Automüügiteavitustes, mis kirjutatud ühel A2-taseme eksamil, moodustavad arvsõnad keskmiselt

10,2% teksti sõnedest. Ülejäänud tekstides kõigub see näitaja 1,2%–1,9% vahel, mis ei erine kuigivõrd B1- ja B2-tasemest. Arvsõnade kasutus eristab eelkõige C1-taset. Tasakaalus korpuses on arvsõnade osakaal 1,6%: enim leidub neid ajakirjandustekstides, ilu- ja teaduskirjanduses aga C1-tasemega üsna sarnasel määral (0,9%).

Sama taseme eksamite lõikes varieerub enim nimi-, omadus- ja asesõnade, eraldi vaadelduna isikuliste asesõnade kasutus. Erinevus on oluline A2–B2-tasemel. A2- ja B1-tasemel varieerub arvsõnade osakaal, B1- ja B2-tasemel tegusõnade, näitavate asesõnade ja alistavate sidesõnade osakaal. Nagu sõnaliigisuhetel põhinevate leksikaalsete tunnuste analüüs esile tõi, sõltub sõnaliikide osakaal mitte üksnes teksti liigist (nt ametlikud ja mitteametlikud kirjad B2-tasemel), vaid ka teemast (vt 3.1.).

3.2.2. Käändsõnatunnused

A2–C1-taset eristavad läbivalt käändevormide arv, käändsõnade ainsuse- ja mitmusevormide ning seestütleva käände vormide osakaal tekstis. Käändevormide arv kasvab tõusvas joones, kuid selle tunnuse põhjal ei saa tasemete vahele selget piiri tõmmata: keskmiselt kasutatakse A2-tasemel 6, B1-tasemel 8, B2-tasemel 9 ja C1-tasemel 11 käänet, samas kõigub A2-taseme tekstide käändevormide arv enamasti vahemikus 5–8 (88,3% tekstidest), B1-tasemel 7–10 (89,2%), B2-tasemel 8–11 (86,7%) ja C1-tasemel 9–12 (92,5%).

A2-tasemelt C1-tasemele liikudes väheneb käändsõnade kasutus ainsusevormis ja sagenevad mitmusevormid – C1-tasemele jõudes lausa neljakordselt. Vahe ainsuse ja mitmuse osakaalus on A2-tasemel enam kui kümnekordne, C1-tasemel vaid ligi kahekordne. Võrreldes Tasakaalus korpuse andmetega (käändsõnade mitmusevormide osakaal 11,5%, ainsusevormide osakaal 41,7%) esinevad B2- ja C1-taseme tekstides käändsõnad mitmuses sagedamini ja ainsuses harvem. Käändsõnade ainsuse- ja mitmusevormide kasutus varieerub kõigil keeleoskustasemetel märkimisväärselt, sõltudes kirjutamisülesandest (vt allpool).

Ainus kääne, mis eristab statistiliselt kõiki järjestikuseid tasemeid, on seestütlev, mille kasutus kokkuvõttes suureneb, ent mitte lineaarselt. B1-tasemel see kasvab, olles ülekaalukalt suurim (3,8%) meediateemalistes tekstides, kus kirjeldatakse infoallikaid, millest teavet saadakse (*ajalehest*, *internetist*, *telerist*, *raadiost*). Ülejäänud B1-taseme eksamikordadel jääb seestütleva vormide keskmine osakaal tekstides vahemikku 0,9%–1,3%, mis ei erine B2-tasemest, kus seestütleva käände osakaal mõnevõrra kahaneb. C1-tasemel kasvab see taas, sarnanedes Tasakaalus korpuse andmetega (2,1%). Niisiis üldistub erinevus ühelt poolt A2- ja B1-taseme ning teisalt B2- ja C1-taseme vahel.

B1–C1-taset eristab nimetava, omastava, seesütleva ja saava käände vormide osakaal. Keeleoskuse arenedes käändekasutus mitmekesistub, rööpselt väheneb nimetava käände kasutus. Suurim muutus toimub B2-tasemel. A2- ja B1-taseme kirjutisi iseloomustab nimetava käände üleüldistamine kontekstidele, kus tuleks kasutada muud käändevormi, nt *annan sulle *kingitus ~ kingituse*, *viis *päevad ~ päeva*, *kasvab *suurem ~ suuremaks*, *asub Astri *keskus ~ keskuses* (grammatiliste käänete asenduste kohta vt Eslon 2010). Nimetava käände ülekasutust põhjustavad ka nimi- ja omadussõna ühildumisvead, nt **hea ~ heas seisukorras*, **helesinine ~ helesinist värvi*, **huvitav ~ huvitaval kontserdil*. C1-taseme tekstides on käändsõnade osakaal nimetavas käändes lähedane Tasakaalus korpusele (17,7%).

Omastava käände kasutus seevastu suureneb. Sagenevad omastavalised täiendid (*osakonna juht*, *maja omanik*), sh mitut eestäiendit sisaldavas nimisõnafraasis (*uue firma töötajad*, *meie asutuse sünnipäev*). B2-tasemel sageneb tagasõnade kasutus, mille laiend koosneb B2- ja C1-taseme kirjutistes tihti mitmest sõnast (*mõne aasta eest*, *huvi seltskondliku elu vastu*). B2- ja eriti C1-tasemel hakatakse omastavat käänet varasemast enam tarvitama täissihitise funktsioonis (*sõlmis lepingu*, *teevad meie elu lihtsamaks*). Tasakaalus korpuses hõlmavad käändsõnad omastavas käändes 13,6%.

Seesütleva kui sagedaima sisekohakääende kasutus väheneb mõnevõrra B2-tasemel ja suureneb C1-tasemel, kuid varieerub samas igal

tasemel teksti teemast tingitult (vt allpool). Seesütleva käände osakaal Tasakaalus korpuses (2,7%) jääb B2-taseme ning B1- ja C1-taseme keskmise osakaalu vahepeale.

Saav kääne, mis A2- ja B1-taseme eksamitekstides valdavalt puudub, esineb B2-tasemel enam kui pooltes kirjutistes ning on C1-tasemel kinnistunud aktiivses kasutuses, olles põhikäänete, alal-, sees- ja seestütleva käände järel kasutatuim käändevorm (*kokkuvõtteks tahan öelda*, *elu on muutunud paremaks*, *soovib edukaks saada*). Tasakaalus korpuses on saavas käändes sõnade osakaal 1,5%.

Kõigil keeleoskustasemetel erineb eksamite lõikes nimetava käände ja ainsusevormide osakaal. A2–B2-tasemel varieerub mitmusevormide ja seesütleva käände kasutus. Ühelgi tasemel ei ilmne olulisi erinevusi teksti käändevormide arvus ja saava käände kasutuses.

Nimetava käände esinemus on seotud nimi- ja omadussõnade sageduse ning millegi või kellegi (nt A2-tasemel müüdava auto, B1-tasemel kohatud koolikaaslaste) kirjeldamisega, kasutades öeldistäidet. Kirjutiste teemast sõltub ka käändsõnade ainsuse ja mitmuse kasutus. Nt A2-tasemel esinevad mitmusevormid sagedamini reisikirjeldustes ja väljasõidule kutsuvates kirjades, ainsusevormid aga automüügiteadetes. B1-tasemel on mitmus sagedam meediaallikaid käsitlevates tekstides, ainsus hiljuti kohatud koolikaaslase omaduste ja tegemiste kirjeldamisel.

3.2.3. Tegusõnatunnused

A2–C1-taseme eksamikirjutisi eristab tegusõna pöördevormide, käskiva kõneviisi ning eitava kõneliigi esinemus. 1. pöörde vormide kasutus muutub järk-järgult harvemaks, 2. pöörde ja käskiva kõneviisi vormid on sagedamad A2- ja B2-taseme kirjades ning 3. pöörde vormid B1-taseme jutustustes ja C1-taseme arutlustes. Eitusvormide osakaal kasvab B2-tasemeni, taandudes mõnevõrra C1-tasemel.

Kuigi tegusõna pöördevormide esinemus sõltub tekstiliigist, on 3. pöörde sagenemine 1. pöörde suhtes ilmselt marker, mis viitab keeleoskustaseme tõusule. Alates B1-tasemest kasutatakse 3. pöörde vorme

sagedamini kui 1. pöörde vorme, suurim erinevus tuleb esile C1-tasemel (vastavalt 10,1% ja 1,8%). See muutus tuleneb kommunikatiivsetest vajadustest: keeleoskuse kasvades nihkub eesmärk oma vahetute kogemuste ja soovide kirjeldamiselt järjest avaramate probleemide arutamisele. 2. pöörde vormid ja käskiv kõneviis on iseloomulikud kirjadele, kus pöördutakse adressaadi poole, et algatada dialoog.

Eitavate lausete kasutamist seostatakse argumenteeriva tekstiga (vt Kasik 2007). Ka keeleõppijailt nõutakse aina enam oma väidete põhjendamist, nii on eitava kõneliigi osakaalu suurenemine ootuspärane. Teisalt sõltub eitava kõne kasutus kirjutise teemast ega sagene järjepidevalt, vaid on suurim B2-tasemel, kus kolme eksami kirjutiste eesmärk oli kirjeldada ja selgitada isiklikke või kodukandi korrashoiuga seotud muresid (*meie linna tänavatel on palju prügi ning ei keegi ei korista seda*; *ma ei ole robot ja ei saa *kõike ~ kõiki töid teha üksi*).

B1–C1-taset eristab tegusõna pöördeliste ja käändeliste vormide, kindla kõneviisi, ainsuse ja umbisikulise tegumoe kasutus. Pöördeliste vormide sagedus väheneb järk-järgult, vastavalt harvenevad ka tegusõnad kindlas kõneviisis ja ainsuses. Samas suureneb nii B2- kui ka C1-tasemel tegusõna käändeliste ja umbisikuliste vormide osakaal.

Umbisikulist tegumoodi ei esine enamikus A2- ja B1-taseme kirjutistes, B2-tasemel aga leidub suuremas osas ja C1-tasemel pea kõigis tekstides. Kõrgematele keeleoskustasemetele omased abstraktsemad teemad tingivad aktiivse tegija taandamist.

Tegusõna käändelistest vormidest on sagedamad *da*- ja *ma*-tegevusnimi, mis enamikus tekstides kasutusel juba A2-tasemel. Samas varieerub nende osakaal A2- ja B1-taseme lõikes märkimisväärselt, kinnistudes B2- ja C1-tasemel. Mineviku kesksõnade kasutus seostub mineviku liitajavormidega, mille osakaal suureneb oluliselt B2-tasemel. *nud*- ja *tud*-kesksõnu leidub vähestes A2- ja B1-taseme tekstides, samas kui B2-tasemel sisaldab neid üle poole ja C1-tasemel valdav osa tekstidest. Laieneb eelkõige täismineviku kasutus: minevikusündmustest ei jutustata üksnes distantsilt, vaid need seotakse hetkeolukorraga (*mul on viimasel ajal tekkinud tunne, et minu tööjõudu kasutatakse ära*).

Tegusõnade vormikasutus erineb keeleoskustasemete piires sõltuvalt teksti esituslaadist ja sisust. Kuni B2-tasemeni varieerub eksamite lõikes tegusõna pöördeliste ja käändeliste vormide, 1. ja 3. pöörde esinemus. A2- ja B1-tasemel varieerub kindla kõneviisi, A2- ja B2-tasemel ainsusevormide ning B1–B2-tasemel eituse kasutus. Ühelgi keeleoskustasemel ei erine oluliselt mineviku kesksõnade ja umbisikulise tegumoe vormide osakaal tekstis. Tegusõnavormide kasutus on kõige ühtlasem C1-taseme eksamite lõikes.

3.2.4. Järeldusi

Järjepidevaid muutusi sõnaliikide ja grammatiliste vormide kasutuses saab seostada tekstide kasvava keerukusega, mõnel juhul järjest pikemate tekstide kirjutamisega. Muutused ei ole aga sageli samasuunalised. Kolmandik morfoloogilistest tunnustest (11 tunnust 30-st), mille alusel erineb olulisel määral vähemalt kaks järjestikust keeleoskustasemete paari, ei sagene ega harvene järk-järgult. Mittelineaarsed muutused on seotud keelekasutuse situatsiooniliste teguritega: a) teksti liigist ja teemast tulenev varieerumine tasemete piires (nt nimi-, omadus-, ase- ja tegusõnad, sees- ja seestütlev kääne); b) eri keeleoskustaseme eksamiülesannete eripära (nt dialoogile suunatud tekstid A2- ja B2-tasemel *vs.* B1-taseme jutustavad ja C1-taseme arutlevad tekstid). Teisalt võib ka samasuunaliselt muutuva tunnuse hajuvus olla suur, sõltudes samuti kirjutamisülesandest (nt käänd- ja tegusõnade arvuvormid).

Ka Vajjala ja Lõo (2014: 123) andmetel osutusid olulisteks A2- ja B1-taset eristavateks tunnusteks hüüd- ja sidesõnade ning tegusõna 2. pöörde vormide kasutus; B2- ja C1-taset piiritlevate tunnustena tulid esile 2. pööre, käskiv kõneviis ja käändevormide arv tekstis. Siinse analüüsi tulemusel selgus, et hüüdsõnade, tegusõna 2. pöörde ja käskiva kõneviisi esinemus sõltuvad kirjutamisülesandest ega peegelda selgelt keelekasutuse arengut lihtsamast keerukamaks. Peamiste B1- ja B2-taset eristavate tunnustena tõid Vajjala ja Lõo välja sõnavara mitmekesisuse mõõdikud. Siinne uurimus kinnitab, et sõnavara mitmekesistub

B2-tasemel rohkem kui B1-tasemel (vt 3.1), kuid neid tasemeid eristavad ka mitmesugused morfoloogilised tunnused.

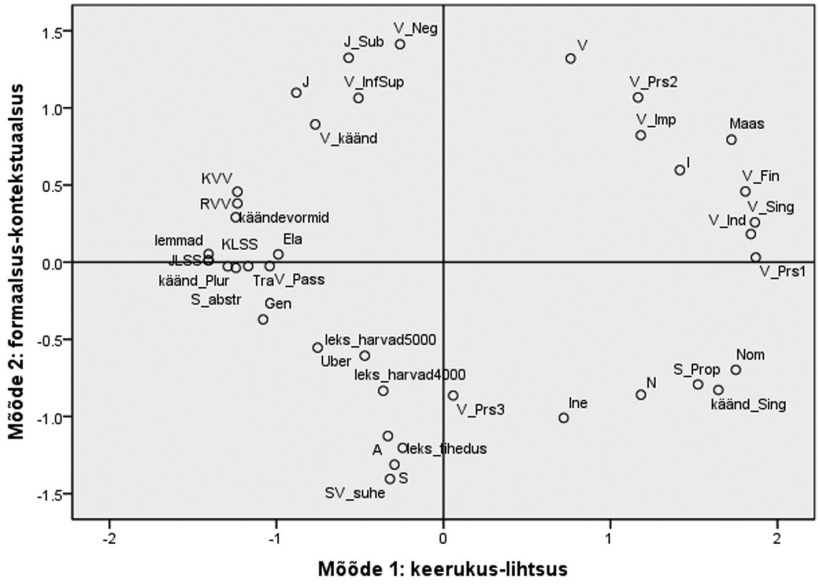
Osa keeleoskustasemeid lineaarselt eristavaid tegusõnatunnuseid on markeerinud C1-taseme arutluste erinevust emakeelekõnelejate arvamustekstidest. Allkivi-Metsoja (2021) andmetel esineb C1-tasemel vähem tegusõna käändevorme ja mineviku liitajavorme, mille kasutus sageneb keeleoskustaseme tõustes, ning samas rohkem kindlat kõneviisi, mis muutub tasemete lõikes harvemaks.

Võrdlus Tasakaalus korpuse statistikaga viitab, et C1-tasemel läheneb sõnaliikide ja käändevormide kasutus eesti kirjakeele mallile. Vilunud keeleõppijalt oodatakse toimetulekut sama nõudlikes keelekasutusolukordades kui haritud emakeelekõnelejalt.

4. Tunnustevahelised seosed

Keeleõppijate loovkirjutiste leksikaalseid ja morfoloogilisi tunnuseid võib seostada nii tekstide lihtsuse-keerukuse tasandiga (muutused A2-tasemelt C1-tasemele liikudes) kui ka situatsioonilistest teguritest mõjutatud kontekstuaalsuse-formaalsuse kontiinumiga. Mitmemõõtmelise skaleerimise meetodil koostatud joonis 1 kujutab vähemalt kahte järjestikust keeleoskustasemete paari eristavate lingvistiliste tunnuste paiknemist neil kahel tasandil. Tunnuste väärtused on standardiseeritud (keskväärtus on 0 ja standardhälve 1). Joonisel 1 on kajastatud 84,6% tunnuste tegelikust varieerumisest.

Joonise 1 horisontaalteljel on paremale koondunud tunnused, mille väärtus on suurem madalamatel keeleoskustasemetel, ja vasakule tunnused, mille väärtus kasvab keeleoskustaseme tõustes. Vertikaaltelje allosas paiknevad tunnused, mis iseloomulikud formaalsemale, sisutihedamale esituslaadile (siinses valimis jutustused ja arutlused), ning ülaosas tunnused, mis omasemad kontekstuaalsemale, enam dialoogilisele suhtlusele suunatud väljendusviisile (siinses valimis kirjad).



Joonis 1. *Keeleoskustasemeid eristavate lingvistiliste tunnuste paiknemine teksti lihtsuse-keerukuse ja formaalsuse-kontekstuaalsuse mõõtmel (kaugusmõõt on eukleidiline kaugus)*

1) Vähesema keerukusega tekste iseloomustavad:

- suurem Maasi indeks (joonisel *Maas*), mis viitab sõnavara vähesele varieerumisele;
- tegusõna ainsuse (*V_Sing*)[3], kindla kõneviisi (*V_Ind*) ja 1. pöörde vormid (*V_Prs1*), pöördelised vormid tervikuna (*V_Fin*);
- käändsõnad nimetavas käändes (*Nom*) ja ainsuses (*käänd_Sing*);
- isikulised asesõnad (*P_Prs*);
- pärisnimede (*S_Prop*), arv- (*N*) ja hüüdsõnade (*I*) suurem osakaal, mis mh seotud tekstide väiksema pikkusega.

[3] Morfoloogiliste tunnuste lühendid on tuletatud Stanza morfomärgenduse metakeelest.

Nimetatud tunnuste koosesinemist kinnitab ka korrelatsioonanalüüs: enamik neist on Pearsoni korrelatsioonikordaja r alusel omavahel vähemalt nõrgalt seotud (vt seose tugevuse tõlgendus 2.2.2.) ehk sõltuvad üksteisest teatud määral. Väga tugevalt on seotud tegusõna pöördeliste vormide ja kindla kõneviisi vormide esinemus, mõlemad on ka keskmise tugevusega korrelatsioonis tegusõna 1. pöörde vormide ja ainsusevormidega. Käändsõnade nimetava käände ja ainsusevormidel on keskmiselt tugev seos pärisnimede ja tegusõna ainsusevormidega, nimetaval käändel samuti arvsõnade ja isikuliste asesõnadega.

2) Keelekasutuse suuremale keerukusele viitavad:

- mitmekesisem sõnavara – suurem lemmade arv tekstis (*lemmad*), varieeruvusindeksite (*JLSS*, *KLSS*, *KVV*, *RVV*, Uberi indeks – *Uber*) kõrgem väärtus;
- nimisõnade suurem abstraktsus (*S_abstr*);
- suurem käändevormide arv tekstis (*käänded)*;
- käändsõnad omastavas (*Gen*), seestütlevas (*Ela*) ja saavas käändes (*Tra*) ning mitmuses (*käänd_Plur*);
- umbisikuline tegumood (*V_Pass*) ja tegusõna käändelised vormid (*V_käänd*), sh *da*- ja *ma*-tegevusnimi (*V_InfSup*);
- küsivad-siduvad (*P_IntRel*) ja näitavad asesõnad (*P_Dem*);
- sidesõnade (*J*), sh alistavate sidesõnade (*J_Sub*) suurem osakaal.

Loetletud leksikaalsed tunnused on enamasti vähemalt keskmise tugevusega seoses. JLSS ja KLSS on väga tugevalt seotud sõnavara absoluutse suurenemise ehk lemmade arvuga tekstis (samas on Uberi indeks lemmade arvuga seotud nõrgalt, Maasi indeks keskmises negatiivses seoses – ka JLSS ja KLSS on nendega seotud keskmiselt). Samuti on väga tugev RVV ja KVV seos JLSS-i ja KLSS-iga. Nimisõnade abstraktsus on tugevalt seotud lemmade arvuga.

Keerukamatele tekstidele iseloomulikud morfoloogilised tunnused on omavahel seotud nõrgalt või keskmiselt. Käändevormide arv on keskmise tugevusega korrelatsioonis käändsõnade saava käände ja mitmusevormide ning küsivate-siduvate asesõnade kasutusega. Saav kääne

korreleerub keskmiselt veel mitmuse ja omastava käändega. Tugevapoolne (üle 0,6) on mineviku kesksõnade ja umbisikulise tegumoe seos, mille tingib *tud*-kesksõna umbisikulistes mineviku liitajavormides.

Sõnavara mitmekesisuse ja abstraktsusega on tugevas või keskmises korrelatsioonis käändevormide arv (tugev on seos indeksitega JLSS ja KLSS), ülejäänud morfoloogilised tunnused seostuvad leksikaalsetega kas keskmiselt või nõrgalt. Uberi indeksiga, mis seotud ka tekstide formaalsusega (vt allpool), on kõigi morfoloogiliste tunnuste korrelatsioon nõrk või olematu.

Lihtsamale ja keerukamale keelekasutusele viitavad tunnused on üldiselt negatiivses korrelatsioonis. Tugev negatiivne seos on ootuspäraselt Uberi ja Maasi indeksi vahel ning käändsõnade mitmusevormidel käänd- ja tegusõnade ainsusevormidega.

3) Formaalsema esituslaadiga seostuvad:

- nimi- (*S*), omadus- (*A*) ja arvsõnade, samuti pärisnimede ohtram kasutus;
- suurem nominaalsus (*SV_suhe*) ja leksikaalne tihedus (*leks_tihedus*);
- sõnavara mitmekesisus (Uberi indeksi alusel) ja ulatus – 4000 ja 5000 sagedama sõna hulka mittekuuluva sõnavara suurem osakaal (*leks_harvad4000*, *leks_harvad5000*);
- tegusõna 3. pöörde vormid (*V_Prs3*);
- seesütlev kääne (*Ine*), sagedamast noomenikasutusest tingitult ka nimetav kääne ning käändsõnad ainsuses.

Eksamitekstide nominaalsus ja leksikaalne tihedus on tugevapoolses korrelatsioonis (ligi 0,7) ning tugevalt seotud nimisõnade osakaaluga. Lisaks on nende tunnustega tugevalt seotud formaalsusindeks F, mis joonisel 1 ei kajastu, kuna eristab vaid B2- ja C1-taset (vt 3.1.). Kõigi nelja tunnuse väärtus on väikseim B2-taseme kirjades ning suurim C1-taseme arutlustes.

Ühelt poolt seostuvad nendega keerukamale keelekasutusele omased tunnused. Sõnavara ulatuse tunnused on keskmiselt tugevas seoses

nimisõnade esinemuse, leksikaalse tiheduse ja F-indeksiga. Kõigi nendega on keskmiselt või nõrgalt seotud Uberi indeks. F-indeksi, leksikaalse tiheduse ja nominaalsusega on keskmiselt või nõrgalt seotud ka omadussõnade kasutus, mis sagedaim C1-tasemel.

Teisalt korreleeruvad F-indeksiga nõrgalt arvsõnad ning leksikaalse tiheduse ja nominaalsusega pärisnimed, mille osakaal on suurim A2-taseme tekstides, seostudes käändsõnade ainsusevormide ja nimetava käände kasutusega (vt punkt 1).

Seesütlev kääne, mille kasutus harveneb B2-tasemeni ja sageneb taas C1-tasemel, on nõrgalt seotud leksikaalse tiheduse, F-indeksi, pärisnimede ja nimisõnade üldise osakaaluga. Tegusõna 3. pööre on sagedaim B1- ja C1-taseme tekstides, mis on keskmiselt formaalsemad kui B2-taseme kirjad, kuid 3. pöörde vormidel puudub siinkirjeldatud tunnustega korrelatsioon. See-eest on neil nõrk negatiivne seos 2. pöörde ja isikuliste asesõnadega, mis iseloomulikud kirjades avalduvale kontekstuaalsusele.

4) Kontekstuaalsemale esituslaadile on omased:

- tegusõnade (*V*), nii nende käändeliste kui ka pöördeliste vormide, eriti 2. pöörde (*V_Prs2*), käskiva kõneviisi (*V_Imp*) ja eituse (*V_Neg*) suurem osakaal (eitus seostub B2-taseme probleemkirjadega);
- hüüd- ja asesõnade, sh isikuliste, näitavate ja küsivate-siduvate asesõnade ohtram kasutus;
- ahtam sõnavara (Maasi indeksi alusel).

Tugev seos ilmneb tegusõna 2. pöörde ja käskiva kõneviisi vormide vahel, hüüdsõnad on kummagi vormidega seotud keskmise tugevusega. Kõik need tunnused on kirjutamisülesannetest tingitult sagedaimad A2- ja B2-tasemel. Maasi indeks korreleerub keskmiselt isikuliste asesõnadega ja nõrgalt tegusõna pöördeliste vormidega, mille osakaal on suurem A2-taseme kirjades. Teisalt on nõrk seos tegusõna eitusvormide ja näitavate asesõnade vahel – mõlemad sagenevad B2-tasemel.

Kontekstuaalsuse-formaalsusega seotud sõnaliikide kasutuses tulevad esile vastandused: ase- ja tegusõnade sagedus on keskmiselt tugevas negatiivses korrelatsioonis nimisõnade sagedusega ning vastavalt keskmises ja nõrgas negatiivses seoses omadussõnade sagedusega. Sidesõnu ei seostata formaalsusega, ent analüüsitud tekstivalimis on alistavad sidesõnad sagedaimad B2-taseme kirjades.

Järeldusi. Keerukuse skaala eri otstesse koonduvad tunnused, mis eristavad A2–C1-taseme eksamitekste üldiselt lineaarselt. Skaala keskosas asetsevad tunnused, mille puhul on muutused tasemete vahel mittelineaarsed, sõltudes peamiselt kirjutamisülesandest. Just need tunnused vastanduvad enamasti formaalsuse mõõtmel, ehkki osa tunnuseid seondub samaaegselt ka keelekasutuse keerukusega (nt Maasi ja Uberi indeks, eri liiki asesõnade osakaal).

Formaalsuse skaala keskele (joonisel 1 horisontaaltelje lähistele) paigutuvad tunnused, mida näib ülesandest tingitud varieerumine vähe mõjutavat ja mida saab lugeda usaldusväärsemaks eri tasemete eksamikirjutiste piiritlemisel. Sellistena kerkivad esile a) leksikaalsetest tunnustest lemmade arv tekstis, indeksid JLSS, KLSS, RVV ja KVV ning nimisõnade abstraktsus; b) morfoloogilistest tunnustest tegusõna 1. pööre, kindel kõneviis, ainsus ja umbisikuline tegumood, käändevormide arv, käändsõnad omastavas, saavas ja seestütlevas käändes ning mitmuses.

Uberi ja Maasi indeksi sõltumine teksti nimisõnarohkusest annab kinnitust, et selgemalt kajastavad õppijate sõnavara rikastumist indeksid JLSS ja KLSS. Viimased on aga tugevalt seotud unikaalsete sõnade arvuga tekstis ega pruugi taseme prognoosimiseks sama hästi sobida, juhul kui tekstide pikkus varieerub tasemete piires rohkem kui eksamikirjutiste puhul.

5. Kokkuvõte

Eksamitekstide automaatöötluse ja statistilise andmeanalüüsi tulemustes sisaldub vastus artikli alguses sõnastatud uurimisküsimustele.

Esiteks vaadeldi, missugused lingvistilised tunnused eristavad eesti keele õppijate A2–C1-taseme kirjutisi ja mil moel: millistel keeleoskustasemetel ja mis suunas toimuvad olulised muutused. Andmetest järeldub, et mitmete teiste keeltega analoogselt saab ka eesti keeles eri tasemete keelekasutust piiritleda nii leksikaalsete kui ka morfoloogiliste tunnuste alusel. Enamik analüüsitud tunnustest (61 tunnust 69-st) eristab järjestikuseid keeleoskustasemeid: osa neist kõiki kolme kõrvuti asetsevate tasemete paari, osa aga kahte või ühte tasemepaari.

Tunnuste dünaamika võib olla lineaarne ja mittelineaarne. Lineaarselt muutuvad tunnused on järjepidevas kasvamis- või kahanemistrendis ja omasemad vastavalt kõrgemate või madalamate tasemete keelekasutusele. Neid tunnuseid saab seostada tekstide suureneva keerukusega ja käsitleda tasemeid eristavate markeritena. Uudsena tõi andmeanalüüs esile üldistatud arvulised hinnangud mh sellele, kui mitmekesine ja abstraktne on eri tasemete eksamikirjutiste sõnavara, mis tasemetel ja millise sagedusega erinevaid sõnaliike, käänd- ja tegusõnavorme kasutatakse ning millal tulevad tarvitusele keerukama keelekasutusega seotud vormid, nt saav kääne, umbisikuline tegumood, mineviku kesksõnad.

Teine küsimus, millele uurimuses vastust otsiti, puudutas kirjutamisülesande mõju tekstide lingvistiliste tunnuste avaldumisele. Mittelineaarsed muutused tasemete vahel tulenevad teksti liigi ja teema varieerumisest kas sama taseme või eri tasemete eksamitel. Erisuunaliselt muutuvad tunnused, eelkõige sõnaliigisagedustel põhinevad leksikaalsed ja morfoloogilised tunnused, seostuvad enamasti kontekstuaalsuse-formaalsuse kontiinumiga. Selles osas vastanduvad kirjad kui dialoogi arendamisele suunatud ja kontekstitundlikumad tekstid ülejäänud eksamikirjutistega – kirjelduste, jutustuste ja arutlustega, mida iseloomustab sisutihedam ja ühemõttelisem esituslaad.

Uurimus aitas välja tuua usaldusväärsemad tunnused, mis seostuvad pigem keelekasutuse keerukuse kui kirjutamisülesandest olenevate teguritega. Neist tunnustest on kavas lähtuda keeleoskustaseme automaathindamiseks statistilisi ennustusmudeleid koostades. Tekstide automaatne klassifitseerimine toob omakorda välja, millised nende tunnuste kombinatsioonid võimaldavad tekstide taset kõige täpsemini prognoosida.

Paindlikuma tasemehindamise jaoks on tarvis leida keeleoskustasemeid eristavad tunnused, mis sõltuvad kirjutamisülesandest ja -olukorrast võimalikult vähe. Selleks tuleb siinse valimi analüüsil saadud tulemusi valideerida mitmekesisema tekstimaterjali alusel, mis hõlmab nii uusi eksamikirjutisi kui ka väljaspool eksamiolukorda kirjutatud tekste, mille taset on hinnanud eksperdid.

Tänusõnad

Artikli valmimist on toetanud Tallinna Ülikooli uuringufond (projektid TF1519 ja TF2019) ning ITL ja HITSA (nüüdne HARNO) Ustus Aguri nimelise stipendiumiga. Tänan töörühma kolleege abi eest tekstide märgenduse parandamisel ja retsensente kasulike soovituste eest.

Võrguviited

Eesti vahekeele korpus. https://evkk.tlu.ee/about (3.10.2021).

EstUD. Estonian Treebank in form of Universal Dependencies. https://github.com/EstSyntax/EstUD (22.8.2021).

Pandas. https://pandas.pydata.org (22.8.2021).

Sampling. https://CRAN.R-project.org/package=sampling (22.8.2021).

Stanza: A Python NLP Package for Many Human Languages. https://stanfordnlp.github.io/stanza (22.8.2021).

Sõnaliikide sagedusloend ning käändsõna grammatiliste kategooriate sagedusloendid Tasakaalus korpuse põhjal. https://cl.ut.ee/ressursid/gram-kat/ (22.8.2021).

Tekstide abstraktsus. http://prog.keeleressursid.ee/abstraktsus/ (22.8.2021).

Kirjandus

Allkivi, Kais 2016. 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-Metsoja, Kais 2021. C1-tasemel eesti keele õppija kirjalik keelekasutus võrdluses emakeelekõnelejaga ['Written language use of C1 learners of Estonian and native speakers in comparison']. – Annekatrin Kaivapalu, Pille Eslon (toim.). Eesti keele oskuse arenemine ja arendamine. Kirjalik õppijakeel. Tallinn: EKSA, 205–231.

Alp, Pilvi, Krista Kerge, Hille Pajupuu 2013. Measuring lexical proficiency in L2 creative writing. – Jozef Colpaert, Mathea Simons, Ann Aerts, Margret Oberhofer (eds.). Language Testing in Europe: Time for a New Framework? Antwerpen: Linguapolis Universiteit Antwerpen, 274–286.

Armstrong, Richard A. 2014. When to use the Bonferroni correction. – Ophthalmic & Physiological Optics 34 (5), 502–508. https://doi.org/10.1111/opo.12131

Arnold, Taylor, Nicolas Ballier, Thomas Gaillat, Paula Lissón 2018. Predicting CEFRL levels in learner English on the basis of metrics and full texts. – Conférence sur l'Apprentissage Automatique, INSA Rouen. https://arxiv.org/pdf/1806.11099.pdf (22.8.2021).

Bartning, Inge, Maisa Martin, Ineke Vedder (eds.) 2010. Communicative Proficiency and Linguistic Development. Intersections between SLA and Language Testing Research. EuroSLA Monograph Series 1. European Second Language Association.

Delacre, Marie, Christophe Leys, Youri L. Mora, Daniël Lakens 2019. Taking parametric assumptions seriously: Arguments for the use of Welch's *F*-test instead of the classical *F*-test in One-Way ANOVA. – International Review of Social Psychology 32 (1), a13. https://doi.org/10.5334/irsp.198

Eslon, Pille 2010. Suundumustest eesti keele grammatiliste käänete kasutuses ['Tendencies in the use of grammatical cases in Estonian']. – Pille Eslon, Katre Õim (toim.). Korpusuuring ja meetodid. Tallinna Ülikooli eesti keele ja kultuuri instituudi toimetised 12. Tallinn: TLÜ EKKI, 7–36.

Eslon, Pille 2021. Eesti keele kasutamine A2- ja B1-taseme tekstides soome- ja venekeelsete õppijate näitel ['Estonian language usage in A2- and B1-level texts of Finnish- and Russian-speaking learners']. – Annekatrin Kaivapalu,

Pille Eslon (toim.). Eesti keele oskuse arenemine ja arendamine. Kirjalik õppijakeel. Tallinn: EKSA, 117–204.

CEFR 2001 = Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: Cambridge University Press.

CEFR 2018 = Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Companion Volume with New Descriptors. Strasbourg: Council of Europe Publishing.

Granger, Sylviane, Martin Wynne 1999. Optimising measures of lexical variation in EFL learner corpora. – John M. Kirk (ed.). Corpora Galore. Amsterdam/ Atlanta: Rodopi, 249–257.

Hancke, Julia 2013. Automatic Prediction of CEFR Proficiency Levels Based on Linguistic Features of Learner Language. MA Thesis. Universität Tübingen.

Harrison, Julia, Fiona Barker (eds.) 2015. English Profile in Practice. Cambridge: Cambridge University Press.

Heylighen, Francis, Jean-Marc Dewaele 2002. Variation in the contextuality of language: An empirical measure. – Foundations of Science 7 (3), 293–340. https://doi.org/10.1023/A:1019661126744

HTM 2007 = Euroopa keeleõppe raamdokument: õppimine, õpetamine, hindamine ['CEFR']. Tartu: Haridus- ja Teadusministeerium, 2007.

Hulstijn, Jan H. 2014 The Common European Framework of Reference for Languages: A challenge for applied linguistics. – International Journal of Applied Linguistics 165 (1), 3–18. https://doi.org/10.1075/itl.165.1.01hul

Kaalep, Heiki-Jaan, Kadri Muischnek 2002. Eesti kirjakeele sagedussõnastik ['Frequency Dictionary of Written Estonian']. Tartu: Tartu Ülikooli Kirjastus.

Kasik, Reet 2007. Sissejuhatus tekstiõpetusse ['Introduction to Textual Study']. Tartu: Tartu Ülikooli Kirjastus.

Kerge, Krista, Hille Pajupuu, Rene Altrov 2007. Tekst, kontekstuaalsus ja kultuur ['Text, contextuality and culture']. – Keel ja Kirjandus 8, 624–637.

Kerge, Krista 2010. Kirjažanrite keeleparameetrid mitme tekstiliigi taustal ['Linguistic parameters of letter genres with regard to oral and written language']. – Emakeele Seltsi aastaraamat 55, 32–62.

Kerge, Krista, Anne Uusen, Halliki Põlda 2014a. Teismeea loovkirjutiste sõnavara ja selle hindamine ['Teenage vocabulary and its assessment in creative writing']. – Eesti Rakenduslingvistika Ühingu aastaraamat 10, 157–175. https://doi.org/10.5128/ERYa10.10

Kerge, Krista, Anne Uusen, Halliki Põlda, Helin Puksand 2014b. Loovkirjutiste süntaksimuutujate areng teismeeas ['Development of syntactic parameters of teenage creative writing']. – Emakeele Seltsi aastaraamat 59, 46–76. https://doi.org/10.3176/esa59.03

Kirt, Riin 2013. Tasakaalus korpusel põhinevad sagedusloendid ja korpuse sõnavara ning "Eesti keele seletava sõnaraamatu" märksõnaloendi võrdlus ['Word frequency lists based on the "Balanced Corpus of Estonian" and selective comparison of corpora frequency lists with keywords from the "Explanatory Dictionary of Estonian"']. Magistritöö. Tartu: Tartu Ülikool.

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']. Dissertations on Humanities 43. Tallinn: Tallinna Ülikool.

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

Kuiken, Folkert, Ineke Vedder 2007. Task complexity and measures of linguistic performance in L2 writing. – International Review of Applied Linguistics in Language Teaching 45 (3), 261–284. https://doi.org/10.1515/iral.2007.012

Lu, Xiaofei 2012. The relationship of lexical richness to the quality of ESL learners' oral narratives. – The Modern Languages Journal 96, 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

McCarthy, Philip M., Scott Jarvis 2007. A theoretical and empirical evaluation of vocd. – Language Testing 24, 459–488. https://doi.org/10.1177/0265532207080767

McCarthy, Philip M., Scott Jarvis 2010. MTLD, Vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. – Behavior Research Methods 42, 381–392. https://doi.org/10.3758/BRM.42.2.381

McEnery, Tony, Richard Xiao, Yukio Tono 2006. Corpus-based Language Studies. An Advanced Resource Book. London/New York: Routledge.

Mikk, Jaan 1979. Õppeteksti keerukus ja õpilaste väljendusoskus ['The complexity of study texts and the students' expression skills']. – Viivi Maanso, Jaan Mikk (toim.). Õppeteksti ja õpilaste väljendusoskuse probleeme. Tallinn: Eesti NSV Pedagoogika Teadusliku Uurimise Instituut, lk 7–12.

Mikk, Jaan, Heiki-Jaan Kaalep, Hiie Asser, Siret Linnas, Merje Songe 2003. Muukeelse kooli 4.–9. klassi eesti keele õpikute tekstianalüüs ['Text analysis of Estonian textbooks for grades 4–9 of the non-Estonian schools']. Tartu: Tartu Ülikool. https://dspace.ut.ee/handle/10062/50110

Mylläri, Taina 2020. Measuring syntactic complexity in learner Finnish. – Apples: Journal of Applied Language Studies 14 (2), 67–92. https://doi.org/10.47862/apples.99134

Pajupuu, Hille, Krista Kerge 2010. Text-types in speech technology and language teaching. – Jorge Luis Bueno Alonso et al. (eds.). Analizar datos > Describir variación. Vigo: Universidade de Vigo, 380–390.

Pajupuu, Hille, Krista Kerge, Pilvi Alp 2009. Sõnavara loomulik rikkus haritud keeleoskaja tekstides ['Natural lexical richness in educated language use']. – Eesti Rakenduslingvistika Ühingu aastaraamat 5, 187–196. https://doi.org/10.5128/ERYa5.12

Pilán, Ildikó 2018. Automatic Proficiency Level Prediction for Intelligent Computer-assisted Language Learning. PhD Thesis. Göteborg: Göteborgs Universitet.

Puksand, Helin, Krista Kerge 2012. Õpikuteksti analüüs kirjaoskuse omandamise kontekstis. – Emakeele Seltsi aastaraamat 57, 162–217. https://doi.org/10.3176/esa57.09

Rowntree, Derek 1981. Statistics without Tears. A Primer for Non-mathematicians. New York: MacMillan Publishing Company.

Rysová, Katerina, Magdaléna Rysová, Jirí Mírovský 2016. Automatic evaluation of surface coherence in L2 texts in Czech. – Proceedings of the 28th international Conference on Computational Linguistics and Speech Processing. Association for Computational Linguistics, 214–228.

Rysová, Katerina, Magdaléna Rysová, Michal Novák, Jirí Mírovský, Eva Hajičová 2019. EVALD: A pioneer application for automated essay scoring in Czech. – The Prague Bulletin of Mathematical Linguistics 113, 9–30. https://doi.org/10.2478/pralin-2019-0004

Siiber, Marten 2018. Rakendus tekstide abstraktsuse hindamiseks ['An application for evaluating the abstractness of texts']. https://dspace.ut.ee/handle/10062/62442 (22.8.2021).

Solovyev, Valery, Marina Solnyshkina, Mariia Andreeva, Andrey Danilov, Radif Zamaletdinov 2020. Text complexity and abstractness: Tools for the Russian language. – Proceedings of the International Conference "Internet and Modern Society", 75–87.

Szügyi, Edit, Sören Etler, Andrew Beaton, Manfred Stede 2019. Automated assessment of language proficiency on German data. – Proceedings of the 15th Conference on Natural Language Processing, 30–39.

Tack, Anaïs, Thomas Francois, Sophie Roekhaut, Cédrick Fairon 2017. Human and automated CEFR-based grading of short answers. – Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications, 169–179.

Treffers-Daller, Jeanine, Patrick Parslow, Shirley Williams 2018. Back to basics: How measures of lexical diversity can help discriminate between CEFR levels. – Applied Linguistics 39 (3), 302–327. https://doi.org/10.1093/applin/amw009

Uiboaed, Kristel 2018. Eestikeelsete stoppsõnade loend ['List of Estonian stop words']. http://www.tekstikaeve.ee/blog/2018-04-18-eestikeelsete-stoppsonade-loend (3.10.2021).

Vajjala, Sowmya, Kaidi Lõo 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), Atlanta, Georgia, June 13 2013. Association for Computational Linguistics, 63–72.

Vajjala, Sowmya, Kaidi Lõo 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, 113–127.

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. Tallinn: Tallinna Ülikool.

Wisniewski, Katrin 2017 Empirical learner language and the levels of the Common European Framework of Reference. – Language Learning 67 (S1), 232–253. https://doi.org/10.1111/lang.12223

Üksik, Tiiu, Jelena Kallas, Kristina Koppel, Katrin Tsepelina, Raili Pool 2021. Estonian as a second language teacher's tools. – Proceedings of the Sixteenth Workshop on Innovative Use of NLP for Building Educational Applications, 130–134.

Yannakoudakis, Helen, Øistein E. Andersen, Ardeshir Geranpayeh, Ted Briscoe, Diane Nicholls 2018. Developing an automated writing placement system for ESL learners. – Applied Measurement in Education 31, 251–267. https://doi.org/10.1080/08957347.2018.1464447

Written Estonian at the levels A2–C1: Comparative automated analysis

KAIS ALLKIVI-METSOJA
Tallinn University

To achieve the communicative purposes of the language proficiency levels defined in the Common European Framework of Reference for Languages (CEFR), a learner needs to acquire lexical and grammatical tools specific to the target language (L2). Yet there is little empirical evidence on language-specific features that mark the development from one level to another. This study aims to determine which linguistic features distinguish A2–C1-level written use of Estonian as L2 and how, i.e., which levels differ significantly and what is the direction of change.

Related research has either focused on observing individual linguistic phenomena at different proficiency levels or automatically predicting the level of writings, while not describing the level-to-level dynamics of analysed features. Hereby, an attempt is made to bridge this gap. Relying on language processing and statistical analysis of the extracted data, two types of features are compared in Estonian proficiency examination writings of distinct levels:

1) lexical features – measures related to various aspects of lexical complexity;
2) morphological features – frequencies of parts of speech (PoS) and grammatical categories of nominals and verbs.

The analysed corpus includes 480 creative writings, each level represented by 120 texts randomly sampled from various examinations. Welch's ANOVA with Bonferroni correction is used to test for significant differences between the proficiency levels, and between the examinations of the same level to detect task-induced variance. For pairwise comparisons of proficiency levels, the Games-Howell post-hoc test is used. Correlation analysis and multidimensional scaling are applied to explore the co-occurrence of the linguistic features in learner texts.

Some of the observed features distinguish the proficiency levels consistently, others distinguish one or two pairs of adjacent levels (A2–B1, B1–B2, B2–C1). The changes from level to level can be linear (i.e., the feature values increase or decrease gradually) or nonlinear (i.e., the changes occur in both directions). Linear changes can be associated with growing complexity in writing and considered more reliable markers of level-to-level progress. Nonlinear changes are caused by variation in text genre and topic, either within examinations of the same level or different levels. These features – mainly features based on PoS frequencies – appear to be related to the contextuality-formality continuum (Heylighen & Dewaele 2002) rather than text complexity.

In conclusion, the study reveals relevant features that can henceforth be used to compile predictive models for automated proficiency assessment, whereas providing a novel insight into writing acquisition of Estonian as L2.

Keywords: natural language processing; CEFR levels; lexical complexity; morphological analysis; written learner language; Estonian

Kais Allkivi-Metsoja
Tallinna Ülikooli digitehnoloogiate instituut
Narva rd 25, 10120 Tallinn, Estonia
kais@tlu.ee

Lisa 1. Welchi F-statistik ja keeleoskustasemete erinevuse olulisus (korrigeeritud olulisusnivoo 0,001)

A2–C1-taseme eksamikirjutiste erinemist lingvistiliste tunnuste lõikes hinnati Welchi ANOVA alusel. Oleva käände vormide osakaalu F-statistik jäi arvutamata, sest olevat käänet ei esine A2-tasemel üheski tekstis ja tunnusel puudub hajuvus. Rühmadevaheline vabadusastmete arv $k - 1 = 3$, kuna võrreldavate rühmade arv $k = 4$.

* Tunnus eristab Gamesi-Howelli järeltesti alusel ühte järjestikuste keeleoskustasemete paari.
** Tunnus ei erista järjestikuseid keeleoskustasemeid.

Tunnuse liik	Tunnus	Welchi F	Rühmadesisene vabadusastmete arv	p-väärtus
Leksikaalsed tunnused	Lemmade arv	1318,9	243,2	< 0,001
	LSS*	104,6	262,1	< 0,001
	JLSS	729,0	259,2	< 0,001
	KLSS	729,0	259,2	< 0,001
	Maasi indeks	75,3	257,5	< 0,001
	Uberi indeks	59,3	255,2	< 0,001
	RVV	226,5	252,4	< 0,001
	KVV	260,8	262,2	< 0,001
	Harvad lemmad (%, mitte 5000 sagedama seas)	67,0	261,6	< 0,001
	Harvad lemmad (%, mitte 4000 sagedama seas)	30,3	261,5	< 0,001
	Harvad lemmad (%, mitte 3000 sagedama seas)*	18,3	260,6	< 0,001
	Harvad lemmad (%, mitte 2000 sagedama seas)*	22,4	263,0	< 0,001
	Harvad lemmad (%, mitte 1000 sagedama seas)*	39,1	260,7	< 0,001
	Leksikaalne tihedus	55,8	263,1	< 0,001
	Nominaalsus (S : V)	25,7	261,8	< 0,001
	F-indeks*	50,9	261,9	< 0,001
	Nimisõnade keskmine abstraktsushinnang	286,3	260,0	< 0,001

Tunnuse liik	Tunnus	Welchi F	Rühmade-sisene vabadus-astmete arv	p-väärtus
Morfoloogilised tunnused: sõnaliikide sagedus	Nimisõnad (%)	43,5	262,1	< 0,001
	Pärisnimed (%)	76,0	246,2	< 0,001
	Omadussõnad (%)	21,9	260,5	< 0,001
	Asesõnad (%)*	78,8	260,1	< 0,001
	Isikulised asesõnad (%)	378,9	243,7	< 0,001
	Enesekohased asesõnad (%)**	11,6	262,2	< 0,001
	Näitavad asesõnad (%)	55,3	264,0	< 0,001
	Küsivad-siduvad asesõnad (%)	78,7	257,5	< 0,001
	Umbmäärased asesõnad (%)*	9,8	253,8	< 0,001
	Arvsõnad (%)	30,0	239,1	< 0,001
	Tegusõnad (%)	9,3	262,0	< 0,001
	Määrsõnad (%)*	11,05	259,7	< 0,001
	Sidesõnad (%)	55,9	257,1	< 0,001
	Rinnastavad sidesõnad (%)*	11,5	258,1	< 0,001
	Alistavad sidesõnad (%)	56,3	260,1	< 0,001
	Kaassõnad (%)	2,4	258,5	0,070
	Tagasõnad (%)*	10,9	259,9	< 0,001
	Eessõnad (%)	5,1	247,8	0,002
Morfoloogilised tunnused: käändsõnatunnused	Käändevormide arv	249,3	264,1	< 0,001
	Käändsõnad nimetavas käändes (%)	133,9	259,6	< 0,001
	Käändsõnad omastavas käändes (%)	82,7	260,0	< 0,001
	Käändsõnad osastavas käändes (%)**	11,7	258,9	< 0,001
	Käändsõnad sisseütlevas käändes (%)*	6,06	256,0	0,001
	Käändsõnad seesütlevas käändes (%)	9,2	257,6	< 0,001
	Käändsõnad seestütlevas käändes (%)	68,3	257,2	< 0,001
	Käändsõnad alaleütlevas käändes (%)*	6,0	257,5	0,001
	Käändsõnad alalütlevas käändes (%)*	7,2	258,0	< 0,001

Tunnuse liik	Tunnus	Welchi F	Rühmade-sisene vabadus-astmete arv	p-väärtus
Morfoloogilised tunnused: käändsõnatunnused	Käändsõnad alaltütlevas käändes (%)*	6,8	257,5	< 0,001
	Käändsõnad saavas käändes (%)	85,3	246,3	< 0,001
	Käändsõnad rajavas käändes (%)*	7,3	248,4	< 0,001
	Käändsõnad ilmaütlevas käändes (%)**	6,8	261,9	< 0,001
	Käändsõnad kaasaütlevas käändes (%)*	7,3	252,4	< 0,001
	Käändsõnad ainsuses (%)	152,1	260,6	< 0,001
	Käändsõnad mitmuses (%)	207,0	262,0	< 0,001
Morfoloogilised tunnused: tegusõnatunnused	Tegusõna pöördelised vormid (%)	88,1	256,6	< 0,001
	Tegusõnad kindlas kõneviisis (%)	91,2	256,7	< 0,001
	Tegusõnad tingivas kõneviisis (%)*	22,4	263,2	< 0,001
	Tegusõnad käskivas kõneviisis (%)	21,8	250,7	< 0,001
	Tegusõnad 1. pöördes (%)	199,1	247,5	< 0,001
	Tegusõnad 2. pöördes (%)	54,3	237,7	< 0,001
	Tegusõnad 3. pöördes (%)	29,5	259,1	< 0,001
	Tegusõnad olevikus (%)*	15,1	253,9	< 0,001
	Tegusõnad lihtminevikus (%)*	12,3	247,8	< 0,001
	Tegusõnad ainsuses (%)	146,7	253,6	< 0,001
	Tegusõnad mitmuses (%)	4,7	254,6	0,003
	Tegusõna eitusvormid (%)	35,0	262,1	< 0,001
	Tegusõnad umbisikulises tegumoes (%)	49,8	256,2	< 0,001
	Tegusõna käändelised vormid (%)	38,2	258,2	< 0,001
	da- ja *ma*-tegevusnimi (%)	18,6	257,3	< 0,001
	Mineviku kesksõnad (%)*	29,6	258,0	< 0,001
	Tegusõnad *des*-vormis (%)*	17,1	260,8	< 0,001