doi:10.5128/ERYa18.03

A2–C1-TASEME EKSAMITEKSTIDE KÄÄNDSÕNAKASUTUS

Kais Allkivi-Metsoja

Ülevaade. Keeleoskustaseme automaatseks hindamiseks on tarvis kindlaks teha mõõdetavad tunnused, mis võimaldavad eri tasemete keelekasutust usaldusväärselt määrata. Siinses artiklis on tähelepanu keskmes eesti keele A2–C1-taseme eksamitekstide käändsõnatunnused. Analüüsitakse käändsõnavormide sagedust ja varieerumist nii summaarselt kui ka eri käändsõnaliikide võrdluses. Tuuakse välja need tunnused, mis on korrelatsioonis keeleoskustasemega ja muutuvad kasvavas või kahanevas suunas, piiritledes järjestikuseid tasemeid. Läbivalt eristavate tunnustena tulevad esile tekstis leidunud käänete arv ning ainsuse ja mitmuse kasutus. Erinevused käändsõnade vormides ilmnevad eelkõige B1–C1-tasemel, olulisemad muutused on seotud saava, nimetava ja omastava käändega.*

Võtmesõnad: keeletöötlus, morfoloogia, keeleoskustasemed, kirjalik õppijakeel, eesti keel

1. Uurimuse lähtekohti

Automaatne keeletöötlus toetab sihtkeele õppimist ja õpetamist kahel viisil: 1) õppijakeele analüüs võimaldab arendada automaatse hindamise ja tagasiside süsteeme, annab ainest keeleoskuse arengu uurimiseks ja õppevara koostamiseks; 2) emakeelekasutuse analüüs võimaldab tuvastada sobiva keerukusega lugemismaterjali ja autentseid keelenäiteid õpiülesannete genereerimiseks (Meurers (2019). Esimesega neist seostub üks keeletehnoloogia põhirakendusi arvutipõhises keeleõppes – kirjalike tekstide automaathindamine, mis pakub õppijatele enesekontrolli võimalust ja toetab õpetajate tööd väiksema või ka suurema kaaluga hindamisolukordades (nt keelekursuste paigutustestid või eksamisooritused).

Tekstide keelelisi tunnuseid, mille kaudu saab prognoosida inimekspertide antud hinnanguid, on püütud kindlaks teha mitmesugustest punktiskaaladest ja keeleoskustasemetest lähtudes (vt ülevaadet Vajjala 2018). Sarnaseid katsetusi on tehtud sihtkeele õppijate tekstide automaatseks hindamiseks Euroopa keeleõppe

* Artikli valmimist on toetanud TLÜ uuringufond (projektid TF 2019, TF 1519) ning ITL-i ja HITSA Ustus Aguri nimeline stipendium.

kuueastmelisel skaalal A1–C2 (Hancke 2013, Tack jt 2017, Pilán 2018, Arnold jt 2018, Yannakoudakis jt 2018, Rysová jt 2019, Szügyi jt 2019, Kerz jt 2021). Selleks arendatud programmid analüüsivad kirjalikku keelekasutust erinevatel tasanditel. Rakendatakse nii pindmist keeletöötlust (sõna- ja lausepiiride tuvastus), mille alusel arvutatakse nn pindtunnuseid (ingl *surface features*), nagu sõnade ja lausete pikkus, aga ka sügavamat lingvistilist analüüsi (teksti lemmatiseerimine, grammatiline analüüs jm). Ka eestikeelsete A2–C1-taseme kirjutiste taset on määratud automaatselt, tuginedes leksikaalsetele ja morfoloogilistele tunnustele (Vajjala, Lõo 2014) või sõnaliigijärjenditele (Kossinski 2018).

Automaatse tasemehindamisega on lähedalt seotud probleemistik, mis puudutab kommunikatiivsete keeleoskustasemete lingvistilise profiili kirjeldamist eri keelte jaoks (vt ülevaadet Wisniewski 2017). Õppijate keelekasutuse automaatanalüüs aitaks funktsionaalsete, keelest sõltumatute tasemekirjelduste kõrval esile tuua tasemetevahelisi erinevusi markeerivad keelespetsiifilised tunnused (ingl *criterial features*, vt Hawkins, Buttery 2010). Keeleoskustaseme automaathindamisele pühendatud uurimused on aga keskendunud taseme prognoosimise tehnilisele küljele (klassifitseerimismeetodi valik, prognooside täpsus), jättes suuresti tähelepanuta, millised arengud õppijate keeleloomes tase-tasemelt ilmnevad ja kuidas neid seletada. Nagu märgivad ka Sowmya Vajjala ja Kaidi Lõo (2014: 124), tuleks statistilised ennustusmudelid siduda eri tasemete keelekasutuse kirjeldusega.

Siinne uurimus püüab seda lünka osaliselt täita. Eesmärk on välja selgitada käändsõnavormide kasutust iseloomustavad tunnused, mis eristavad eesti keele A2–C1-taseme kirjalikke tekste ja mida võiks edaspidi keeleoskustaseme automaathindamisel arvesse võtta. Kõrvutatakse käändsõnavormide sagedust ja varieerumist summaarselt (arvestades nii nimi-, omadus-, ase- kui ka arvsõnade kasutust) ning eraldi nimi-, omadus- ja asesõnade lõikes (arvsõnade vormikasutust väikese esinemuse tõttu ei vaadelda). Otsitakse vastust kahele küsimusele:

1. Millised käändsõnatunnused eristavad järjestikuseid keeleoskustasemeid, muutudes samasuunaliselt?
2. Millised tunnused on keeleoskustasemega tugevamalt seotud?

Analüüs põhineb eesti keele tasemeeksamite loovtekstide automaattöötluse andmetel. Valimisse kuulub 120 teksti igalt keeleoskustasemelt. Sama materjali toel on kirjeldatud leksikaalsete ja morfoloogiliste tunnuste (sõnaliikide, käänd- ja tegusõnavormide kasutus) olulisemaid tasemetevahelisi erinevusi, kuid käändsõnavormide kasutust vaadeldi üksnes kokkuvõtlikult, käändsõnaliike eristamata (Allkivi-Metsoja 2021). Vormide sagedused arvutati osakaaluna teksti sõnede suhtes, nagu on varem teinud Vajjala ja Lõo (2014). Uurimuses ilmnes, et eri liiki käändsõnade esinemus sõltub kirjutamisülesandest (teksti liigist ja teemast), varieerudes nii keeleoskustasemete võrdluses kui ka tasemete piires. Nii annab iga taseme sõnaliigispetsiifiliste vormieelistuste kohta usaldusväärsemat teavet vormisageduste leidmine vastava sõnaliigi kasutusjuhtude suhtes. Käändsõnavormide summaarsed osakaalud on siin leitud teksti käändsõnade arvu suhtes (sarnaselt on grammatiliste vormide osakaalusid arvutanud nt Hancke 2013, Pilan 2018 ning Szügyi jt 2019).

Olulisemate keeleoskustasemeid eristavate käändsõnatunnuste määramisel toetutakse korrelatsioonanalüüsile ja tasemetevaheliste erinevuste statistilisele olulisusele. Taseme tõustes ilmnevaid vormisageduse muutusi aitab tõlgendada võrdlus eesti kirjakeele andmetega. Kuna kvantitatiivsed mõõdikud ei väljenda sageli

keelekasutuse järjepidevat arengut, eristades üksteisest vaid teatud keeleoskustasemeid (vt nt Mylläri 2020), siis antakse rööpselt kasutusnäidetel põhinev kvalitatiivne ülevaade käändsõnavormide funktsionaalsest varieerumisest eri tasemetel.

Eesti õppijakeele uurimustes on keeleoskustasemega seotult tähelepanu osutatud tegusõnakasutusele (nt Allkivi 2016, Kitsnik 2018) ning sõnaliigijärgnevustele ja nende piires esinevale grammatilisele-leksikaalsele varieerumisele (nt Voolaid 2018, Eslon, Kaivapalu 2020). Analüüsitud on ka käändekasutust, ent mitte keeleoskustasemete kaupa (nt Eslon 2008, Eslon, Matsak 2009). Kitsamalt on uuritud sihitise käändevalikut (nt Pool 2007, Ehala 2009, Eslon 2011).

Praegu töötatakse eesti keele jaoks välja keeleoskustasemete sõnavara- ja grammatikapädevuse kirjeldusi, milles lähtutakse olemasoleva õppevara ja eesti keele ühendkorpuse kõrval ka noorte ja täiskasvanud õppijate keelekasutusnäidetest (Kallas jt 2020, Üksik jt 2021). Grammatikapädevuse kirjeldustes kajastub, milliseid muutevorme ja mis funktsioonides õppija teatud tasemel eeldatavasti tunneb. Siinne statistiline analüüs annab teavet selle kohta, kui sageli õppija oma koostatud tekstides tegelikult üht või teist vormi kasutab ning kuidas nende sagedus keeleoskustasemeti muutub. Kuna silmas on peetud automaathindamise vajadusi, on siinne vaatepunkt õppijakeskne, seades fookusesse keeleproduktsiooni. Tegemist on korpusest tuleneva (ingl *corpus-driven*) uurimusega, mis toob tasemeomase kirjaliku keelekasutuse tunnusjooned esile objektiivselt, sõltumata keeleõppe ekspertide kogemusel või õppematerjalidel põhinevatest ootustest.

2. Tekstimaterjal ja uurimismeetodid

2.1. Eksamitekstide valim

Õppijakeelekorpusi saab keeleoskustasemetele omase keelekasutuse empiiriliseks analüüsiks kasutada eeldusel, et määratud on tekstide tase vastavalt skaalale. Kuigi lähtutud on ka kooliastmest või keeleõpperühma tasemest, peetakse usaldusväärseimaks kriteeriumiks siiski eksperthinnanguid (Wisniewski 2017: 235–236). Aluseks võidakse võtta eksamil antud hinnang, et tekst vastab testitava taseme nõuetele (nt Cambridge Learner Corpus, vt Hawkins, Buttery 2010). Samuti kasutatakse eksamiväliselt kirjutatud tekstide spetsiaalset hindamist (nt rootsi õppijakeele korpus SweLL, vt Volodina jt 2016).

Siinse uurimuse tarvis on analüüsitud eesti keele tasemeeksamite loovkirjutisi, mis on loetud eksami tasemele vastavaks juhul, kui kirjutamisosa on hinnatud vähemalt rahuldavaks (60% punktidest).[1] Tekstid ja autoreid iseloomustavad anonüümsed metaandmed on saadud SA Innovelt. Käsikirjalised tekstid on ümber trükitud algset kirjaviisi säilitades, pseudonüümistatud ja lisatud Eesti vahekeele ehk õppijakeele korpuse[2] andmebaasi (eksaminandide tausta kohta vt lähemalt Allkivi-Metsoja 2021: 18).

Valim koosneb 480 tekstist, mis jagunevad võrdselt A2, B1-, B2- ja C1-taseme vahel. Keskmine teksti pikkus on vastavalt 46,8 sõnet (standardhälve 12,0), 110,4 sõnet (SH 21,1), 166,1 sõnet (SH 29,2) ja 259,8 sõnet (SH 44,25). Tekstid on pärit 2018. aasta eksamitelt, C1-taseme puhul ka 2017. aastast. A2-, B1- ja B2-taseme

[1] Eksaminande, kes said vähemalt rahuldava kirjutamisoskuse hinnangu, oli 2018. aastal eksami positiivselt sooritanute hulgas A2-tasemel 84%, B1-tasemel 67% ja B2-tasemel 73%. C1-tasemel oli neid 2018. aastal 26% ja 2017. aastal 34% (SA Innove andmed).

[2] https://evkk.tlu.ee/about (12.5.2022).

eksamitelt on omakorda valimisse võetud võrdne arv juhuslikke tekste. C1-taseme 2017.–2018. aasta eksamite põhjal on koostatud ühine juhuvalim, kus kõigi eksamite tekstid on proportsionaalselt esindatud. Tekstide valikut illustreerib tabel 1, mis annab ühtlasi ülevaate eksamiülesannetest.

Tabel 1. Valimi koostamine ja eksamiülesanded

Tase	Eksam	Sobivaid tekste	Valitud tekste	Teksti liik ja teema
A2	2018 I	114	30	teade: kutse väljasõidule
	2018 II	190	30	teade: automüügi kuulutus
	2018 III	96	30	teade: katkine kodumasin
	2018 IV	90	30	kirjeldus: viimane reis
B1	2018 I	165	30	jutustus: lemmikloomade pidamine
	2018 II	278	30	isiklik kiri: kohtumine koolikaaslasega
	2018 III	144	30	jutustus: kontserdil käik
	2018 IV	124	30	jutustus: meediakasutus
B2	2018 I	102	30	isiklik kiri: tööprobleem
	2018 II	97	30	ametlik kiri: kaebus omavalitsusele
	2018 III	55	30	isiklik kiri: raamatute äraandmine
	2018 IV	89	30	ametlik kiri: töökoha sünnipäev
C1	2017 I	22	17	arutlus: õnnetusjuhtumite ennetamine
	2017 II	28	22	arutlus: globaliseerumine
	2017 III	18	15	arutlus: ajakirjanduse mõju
	2017 IV	24	19	arutlus: linnastumine
	2018 I	18	13	arutlus: ühiskondlik sallivus
	2018 II	21	15	arutlus: digipädevused
	2018 III	7	5	arutlus: kodanikuühendused
	2018 IV	18	14	arutlus: rahvatervis, töötervishoid

2.2. Andmetöötlus ja -analüüs

Tekstide morfoloogiliseks märgendamiseks on kasutatud Pythoni keeletöötluspaketti Stanza[3] (Qi jt 2020). Automaatmärgendus on käsitsi kontrollitud ja parandatud (vt ka Allkivi-Metsoja 2021: 19). Märgenduse alusel on arvutatud 70 käändsõnavormide kasutust iseloomustavat arvtunnust: 17 summaarset, nimi- ja asesõnatunnust ning 19 omadussõnatunnust. Käändsõnade lõikes tervikuna ja sõnaliikide kaupa on leitud käändevormide arv, samuti ainsuse- ja mitmusevormide ning 14 käände vormide osakaal vastavalt kõigi käändsõnade ja konkreetse käändsõnaliigi tekstisageduse suhtes (erandina jäi kõrvale omadussõnade rajav kääne, kuna see ei esinenud üheski tekstis). Sisseütleva käände pikka ja lühikest vormi vaadeldakse väikese ja sarnase kasutussageduse tõttu koos. Omadussõnadel on arvutatud alg-, kesk- ja ülivõrde vormide osakaal. Andmestik on koostatud Pythoni andmeanalüüsi paketiga

[3] https://stanfordnlp.github.io/stanza (12.5.2022).

Pandas[4] (McKinney 2010), edasine statistiline analüüs on tehtud tarkvaraga SPSS Statistics (versioon 26.0).

Keeleoskustasemete erinevuse statistilist olulisust on hinnatud Welchi ANOVA ehk parandatud F-statistiku alusel, mis ei eelda erinevalt traditsioonilisest dispersioonanalüüsist, et tunnuste hajuvus võrreldavates rühmades on sarnane. Käändsõnakasutuse tunnused varieeruvad keeleoskustasemeti sageli erineval määral. Samuti on Welchi ANOVA suhteliselt vähetundlik selle suhtes, kui tunnuste väärtuste jaotus ei vasta igas rühmas normaaljaotusele (vt Delacre jt 2019). See tingimus pole täidetud tunnustel, mille jaotus kaldub ühel või mitmel keeleoskustasemel väiksemate väärtuste suunas. Sellised on grammatilised vormid, mille esinemus püsib madalamatel tasemetel (A2, B1) või läbivalt väga väike.

Welchi F-statistik jäi arvutamata käändevormidel, mida vähemalt ühe taseme tekstides ei esinenud (seega puudus tunnusel hajuvus) – selliseid tunnuseid oli 14. Olulisustesti rakendati 56 tunnusele (vt lisa 1), ent ainsuse- ja mitmusevormide osakaalusid võib käsitleda ühe ja sama tunnuse variantidena, sest nende summa on alati 100%. Niisiis kontrolliti kokkuvõttes 52 käändsõnatunnuse olulisust keeleoskustasemete eristamisel. Selleks et vähendada korduvast testimisest tulenevate juhuslike eksimuste tõenäosust, korrigeeriti valitud olulisusnivood 0,05 Bonferroni meetodil (vt Armstrong 2014): $0{,}05 \div 52 = 0{,}001$.

Kui ilmnes statistiliselt oluline erinevus (F-statistiku olulisuse tõenäosus $p \leq 0{,}001$), siis tehti Gamesi-Howelli järeltestiga kindlaks, milliste tasemete tekste tunnus piiritleb. Meetod võtab arvesse korduvat paarikaupa testimist ja võimaldab vastavalt olulisusnivoole arvutada parandatud p-väärtuse. Tasemepaari erinevus loeti oluliseks, kui $p \leq 0{,}05$. Welchi ANOVA tulemustest lähtudes ei tehtud viie tunnuse järeltesti (vt lisa 1). Gamesi-Howelli testi rakendati erandina kahele tunnusele, mille jaoks F-statistikut arvutada ei saanud: kuna saava käände summaarne sagedus ja omadussõna saava käände vormide sagedus suurenevad järjepidevalt, võrreldi paarikaupa ka nimi- ja asesõnade saava käände kasutust eri keeleoskustasemetel.

Käändsõnatunnuste seost teksti tasemega hinnatakse Spearmani astakkorrelatsioonikordaja ρ (roo) alusel, mis sobib seoste mõõtmiseks nii arv- kui ka järjestustunnuste (nagu keeleoskustase) vahel ja mille väärtust tõlgendatakse sarnaselt Pearsoni lineaarse korrelatsioonikordajaga. Korrelatsioon keeleoskustasemega sõltub sellest, kas tunnuse väärtus muutub samas suunas või mitte ja kas tunnus eristab kõiki järjestikuseid tasemeid või osasid neist. Negatiivne seos osutab, et keeleoskustaseme tõustes tunnuse väärtus kahaneb.

Samasuunaliste muutuste esiletoomisel keskendutakse kõrvuti asetsevate keeleoskustasemete võrdlusele, kuid arvestatakse ka erinevusi ülejäänud tasemete vahel: kui tunnuse väärtus oluliselt kasvab või kahaneb, siis peab see olulisel määral erinema kõigist eelnevatest tasemetest, mitte ainult vahetult eelnevast tasemest. Samuti ei loeta oluliseks tunnuseid, mis eristavad küll kaht järjestikust taset, kuid pole üldiselt teksti tasemega korrelatsioonis (põhjuseks võib olla näiteks tunnuse suur varieerumine tasemete piires).

Võrdlusainest kirjakeele tendentsidega pakuvad Tasakaalus korpuse (TK) käändsõna grammatiliste kategooriate sagedusloendid[5], kuid tuleb arvestada, et need põhinevad korpusel tervikuna, mitte eraldiseisvatel tekstidel. Käändsõnavormide funktsionaalset varieerumist kirjeldatakse kvalitatiivselt, tuginedes kasutusnäidete konteksti analüüsile.

[4] https://doi.org/10.5281/zenodo.3509134 (12.5.2022).
[5] https://cl.ut.ee/ressursid/gram-kat (12.5.2022).

3. Keeleoskustasemeid eristavad käändsõnatunnused

A2–C1-taseme eksamitekstide võrdleva statistilise analüüsi tulemused on kokku võetud tabelites 2–4 (vt ka lisa 1). Tabelis 2 on välja toodud käändsõnakasutuse 26 tunnust, mille absoluutarvuline seos keeleoskustasemega on Spearmani ρ alusel > 0,4. Seda võib käsitleda keskmise tugevusega seosena, seost alates 0,7 loetakse üldiselt tugevaks (vt Guilford 1973, Rowntree 1981).

Tabel 2. Keeleoskustasemega tugevamalt seotud tunnused

Jrk	Tunnus	Spearmani ρ
1.	Asesõna käändevormide arv	0,794
2.	Nimisõna käändevormide arv	0,784
3.	Summaarne käändevormide arv	0,780
4.	Omadussõna käändevormide arv	0,756
5.	Käändsõnad ainsuses/mitmuses	–0,722/0,722
6.	Nimisõnad ainsuses/mitmuses	–0,720/0,720
7.	Käändsõnad saavas käändes	0,709
8.	Nimisõnad saavas käändes	0,649
9.	Käändsõnad nimetavas käändes	–0,627
10.	Nimisõnad omastavas käändes	0,572
11.	Käändsõnad omastavas käändes	0,566
12.	Omadussõnad ainsuses/mitmuses	–0,564/0,564
13.	Asesõnad seestütlevas käändes	0562
14.	Omadussõnad omastavas käändes	0,559
15.	Omadussõnad saavas käändes	0,518
16.	Asesõnad nimetavas käändes	–0,517
17.	Omadussõnad nimetavas käändes	–0,515
18.	Asesõnad ainsus/mitmuses	–0,512/0,512
19.	Käändsõnad seestütlevas käändes	0,492
20.	Asesõnad saavas käändes	0,474
21.	Asesõnad seesütlevas käändes	0,464
22.	Omadussõnad keskvõrdes	0,443
23.	Omadussõnad algvõrdes	–0,442
24.	Omadussõnad seestütlevas käändes	0,424
25.–26.	Nimisõnad seestütlevas käändes	0,409
25.–26.	Omadussõnad seesütlevas käändes	0,409

Ülejäänud kaks tabelit kajastavad samasuunalisi muutusi käändsõnavormide esinemises. Tabel 3 annab ülevaate käändekasutuse varieerumisest eri keeleoskustasemete tekstides. Tabelis 4 on esitatud käändsõnavormid, mille sagedus kasvab või kahaneb keeleoskustaseme tõustes. Tähistatud on statistiliselt olulised erinevused järjestikuste tasemete vahel ja võrdluseks lisatud TK andmed. Keskväärtused on

esitatud usaldusvahemikena ehk tegelike keskväärtuste prognoositavate vahemikena (tõenäosusega 95%).

Tabelites 3 ja 4 esile toodud 33 tunnust korreleeruvad keeleoskustasemega vähemalt tugevusega 0,2 ja on enamjaolt esindatud ka tabelis 2 (kõigil juhtudel on seos statistiliselt oluline korrigeeritud olulisusnivool 0,001). Lisaks tuli teksti tasemega tugevamalt seotud tunnuste seas esile kaks erisuunaliselt muutuvat tunnust – seestütleva käände osakaal tervikuna ja nimisõnadel eraldi – ning üks kasvava väärtusega tunnus, mis järjestikuseid tasemeid ei erista: seestütleva käände osakaal omadussõnadel. Järgnevalt kirjeldatakse nende oluliseks osutunud tunnuste põhjal käändsõnakasutuse tendentse läbi keeleoskustasemete.

Tabel 3. Käändekasutuse varieerumine keeleoskustasemeti (eristab tasemeid läbivalt)

Tunnus	**Keskväärtuse usaldusvahemik (95%) ja standardhälve**			
	A2	**B1**	**B2**	**C1**
Summaarne käändevormide arv	6,3 ± 0,2 1,3	8,2 ± 0,2 1,3	9,2 ± 0,3 1,4	10,6 ± 0,2 1,2
Nimisõna käändevormide arv	5,2 ± 0,2 1,2	7,2 ± 0,3 1,4	8,0 ± 0,3 1,4	10,0 ± 0,2 1,3
Omadussõna käändevormide arv	1,4 ± 0,1 0,8	2,1 ± 0,2 1,0	2,8 ± 0,2 1,2	5,4 ± 0,3 1,5
Asesõna käändevormide arv	3,2 ± 0,2 1,2	4,7 ± 0,2 1,0	6,1 ± 0,2 1,4	7,3 ± 0,2 1,3

Tabel 4. Samasuunalised muutused käändsõnavormide kasutuses, erinevused keeleoskustasemete lõikes (olulisusnivoo 0,05) ja võrdlus Tasakaalus korpuse (TK) andmetega

Tunnus	**Keskväärtuse usaldusvahemik (95%) ja standardhälve**				**Olulised erinevused tasemete vahel**			**TK andmed**
	A2	**B1**	**B2**	**C1**	**A2–B1**	**B1–B2**	**B2–C1**	
Summaarsed tunnused (%)								
Nimetav	51,6 ± 2,0 10,7	52,4 ± 1,5 8,2	42,2 ± 1,5 8,0	34,4 ± 1,2 6,8		✓	✓	33,2
Omastav	11,9 ± 1,4 7,6	11,1 ± 0,9 5,0	18,7 ± 1,3 7,2	22,1 ± 1,0 5,6		✓	✓	25,5
Osastav	10,6 ± 1,4 7,4	11,5 ± 1,2 6,4	14,7 ± 1,2 6,8	15,3 ± 0,9 4,7		✓		14,6
Saav	0,1 ± 0,1 0,7	0,4 ± 0,2 0,8	1,4 ± 0,3 1,8	3,6 ± 0,4 2,4	(✓)*	✓	✓	2,9
Ainsus	91,9 ± 1,5 8,3	83,0 ± 1,7 9,3	75,6 ± 2,3 12,4	65,7 ± 1,4 7,5	✓	✓	✓	77,1
Mitmus	8,1 ± 1,5 8,3	17,0 ± 1,7 9,3	24,4 ± 2,3 12,4	34,3 ± 1,4 7,5	✓	✓	✓	22,9
Nimisõnatunnused (%)								
Nimetav	42,3 ± 3,1 17,0	41,4 ± 2,1 11,6	35,9 ± 1,8 10,1	29,4 ± 1,2 6,6		✓	✓	28,0
Omastav	9,6 ± 1,6 8,7	9,6 ± 1,3 7,2	16,8 ± 1,6 8,7	22,6 ± 1,2 6,3		✓	✓	26,7

Tunnus	Keskväärtuse usaldusvahemik (95%) ja standardhälve				Olulised erinevused tasemete vahel			TK andmed
	A2	B1	B2	C1	A2–B1	B1–B2	B2–C1	
Osastav	12,6 ± 1,9 10,6	17,0 ± 1,7 9,5	19,9 ± 1,8 10,1	17,2 ± 1,0 5,7	✓			15,0
Saav	0,0 0,0	0,4 ± 0,2 1,2	1,5 ± 0,5 2,5	3,1 ± 0,4 2,3	(✓)	✓	✓	2,9
Ainsus	94,9 ± 1,2 6,8	83,8 ± 2,2 11,9	77,7 ± 2,5 13,5	66,2 ± 1,6 8,6	✓	✓	✓	77,5
Mitmus	5,1 ± 1,2 6,8	16,2 ± 2,2 11,9	22,3 ± 2,5 13,5	33,8 ± 1,6 8,6	✓	✓	✓	22,5
Omadussõnatunnused (%)								
Nimetav	80,7 ± 5,0 25,8	77,3 ± 3,7 20,0	68,3 ± 4,0 21,7	49,4 ± 2,8 15,2		✓	✓	41,2
Omastav	2,9 ± 2,5 12,8	4,6 ± 1,7 9,1	10,9 ± 2,2 11,9	14,5 ± 1,7 9,1		✓	✓	24,6
Seesütlev	0,2[0,0; 0,6]** 1,9	1,5 ± 0,9 5,0	1,8 ± 1,0 5,4	4,4 ± 1,1 6,2			✓	3,9
Saav	0,2[0,0; 0,5] 1,6	1,0 ± 0,7 3,6	2,8 ± 1,2 6,8	6,7 ± 1,3 7,0			✓	5,5
Ainsus	94,6 ± 2,8 14,5	83,7 ± 3,5 19,2	75,9 ± 4,3 23,5	67,4 ± 2,3 12,3	✓	✓	✓	75,1
Mitmus	5,4 ± 2,8 14,5	16,3 ± 3,5 19,2	24,1 ± 4,3 23,5	32,6 ± 2,3 12,3	✓	✓	✓	24,9
Algvõrre	97,8 ± 1,6 8,3	93,1 ± 2,2 12,2	92,5 ± 1,9 10,4	89,3 ± 1,5 8,4	✓			90,9
Keskvõrre	1,8 ± 1,4 7,1	6,6 ± 2,2 11,7	6,6 ± 1,8 9,6	10,2 ± 1,5 8,3	✓		✓	7,8
Asesõnatunnused (%)								
Nimetav	57,7 ± 3,8 20,4	58,2 ± 2,1 11,6	42,4 ± 1,9 10,1	38,0 ± 1,9 10,3		✓	✓	43,2
Omastav	19,5 ± 2,9 15,5	15,6 ± 1,4 7,8	23,7 ± 1,9 10,6	26,0 ± 1,9 10,1		✓		22,4
Osastav	6,1 ± 1,8 9,6	5,1 ± 1,1 6,0	10,9 ± 1,4 7,7	11,6 ± 1,2 6,5		✓		15,0
Seesütlev	0,1[0,0; 0,3] 1,0	0,5 ± 0,3 1,5	1,5 ± 0,5 2,5	2,7 ± 0,6 3,3		(✓)	✓	2,2
Seestütlev	0,2[0,0; 0,5] 1,7	0,8 ± 0,4 1,9	1,6 ± 0,4 2,4	4,5 ± 0,7 4,0		✓	✓	3,3
Saav	0,0 0,0	0,1 ± 0,1 0,6	0,6 ± 0,3 1,6	2,4 ± 0,5 3,0		(✓)	✓	0,8
Kaasaütlev	1,3 ± 0,7 3,8	1,7 ± 0,6 3,1	1,9 ± 0,5 2,8	3,2 ± 0,6 3,2			✓	1,5
Ainsus	84,6 ± 3,3 18,1	81,5 ± 2,3 12,4	69,5 ± 3,0 16,6	62,9 ± 2,3 12,5		✓	✓	82,6
Mitmus	15,4 ± 3,3 18,1	18,5 ± 2,3 12,4	30,5 ± 3,0 16,6	37,1 ± 2,3 12,5		✓	✓	17,4

* Sulgudesse on märgitud statistiliselt olulised, ent tegelikkuses mitteolulised erinevused, mille korral esineb käändevorm osakaalu suurenemist hoolimata väga vähestes tekstides.

** Kuna osakaal saab olla üksnes positiivne arv, on mõningad usaldusvahemikud ebasümmeetrilised. Need on märgitud kandilistesse sulgudesse.

3.1. Käändevormide kasutus

Käändevormide varieerumine eristab A2–C1-taset järjepidevalt. Igal järgneval tasemel suureneb statistiliselt olulisel määral nii tekstis kasutatud käänete arv summaarselt kui ka nimi-, omadus- ja asesõnade käändevormide arv eraldi võttes (vt tabel 3). Need neli tunnust on ka keeleoskustasemega tugevaimalt seotud (vt tabel 2). Kõige rikkalikum on käänete varieerumine nimisõnadel. Nimetava, omastava ja osastava kõrval hakatakse esimesena kasutama sees- ja alalütlevat käänet (eeskätt koha- ja ajamäärustena), harvem sisse- ja alaleütlevat ning kaasaütlevat. B1-tasemel muutub olulisemaks ka seestütlev ja B2–C1-tasemel saav kääne. Ülejäänud käändevormid moodustavad läbivalt alla 1% nimisõnakasutustest.

Omadussõnu kasutatakse A2-tasemel enamasti vaid ühes käändevormis – tavaliselt nimetavas, harvem osastavas (*sul on vaba aega*). Sagedamate käänete hulka kuuluvad veel omastav ja alalütlev (*aasta algab uue töökoha otsinguga*, *aitavad raskel hetkel*), C1-tasemel ka saav, sees- ja seestütlev kääne (*maailm on nii väikseks läinud*, *uued elanikud ei osale kohalikus elus*, *rääkisime väga olulistest probleemidest*).

3.2. Ainsuse- ja mitmusevormide kasutus

Järjest sagedamini tarvitatakse käändsõnade mitmusevorme, selle arvelt väheneb ainsusevormide osakaal. Tasemete lõikes suureneb pidevalt nimi- ja omadussõnade kasutus mitmuses, asesõnade mitmusevormide osakaal kasvab B2- ja C1-tasemel. Käändsõnade summaarne ning samuti nimisõnade jaotumine ainsuse- ja mitmusevormide vahel on keeleoskustasemega tugevalt seotud, omadus- ja asesõnade arvuvormide kasutus korreleerub teksti tasemega nõrgemini. Omadussõnadel on ainsuse ja mitmuse sageduse varieerumine suurim.

Keerukamate grammatiliste vormide, k.a mitmuse kasutamine sageneb kooskõlas kommunikatiivsete vajadustega. B1-taseme kirjutiste teemadering, nt lemmikloomade pidamine ja meelepärased infoallikad (*kassid on targemad kui koerad*, *Postimehel on alati värsked uudised ja huvitavad teemad*), eeldab nimi- ja omadussõnade avaramat kasutust mitmuses kui A2-taseme kirjutamisülesanded. B2-taseme kirjades kasvab enim asesõna mitmusevormide osakaal. Näiteks jutustatakse *meie*-vormis endast ja oma töökaaslastest või naabritest (*meil on üsna palju kliente*, *meie maja elanikud ei ole selle firma tööga rahul*), pakutakse tuttavatele kodus seisma jäänud raamatuid ja ajakirju (*tahan need teile ära anda*) ning kasutatakse ametikirjadele omast viisakus-*Teie* vormi (*pöördun Teie poole kaebusega*).

Järjekordse nihke toob kaasa üleminek isiklikelt teemadelt üldisematele. C1-taseme arvamustekstides arutletakse ühiskondlike probleemide ja nähtuste üle, nagu digiseadmete kasutus või sisse- ja väljaränne, millega seoses on loomulik viidata osalistele mitmusevormis (*inimesi ümbritsevad masinad ja robotid*, *noored haritud inimesed kolivad suurematesse linnadesse*). Sagedaim isikuline asesõna on *meie*, millega tähistatakse enamasti inimkonda tervikuna või Eesti elanikke, vahel ka kodukoha elanikke vm kitsamat inimrühma (*Meie maailm areneb suure kiirusega*, *me elame e-riigis*, *Meie linngi on viimastel aastatel jõudsalt kasvanud*).

C1-taseme kirjutistes on käändsõnade mitmusevormide kasutus suurem kui TK-s, mille statistikaga sarnaneb enim ainsuse- ja mitmusevormide jaotumine B2-tasemel, asesõnade puhul aga B1-tasemel.

3.3. Grammatiliste käänete kasutus

Sarnaselt ainsusega kahaneb ka nimetava käände osakaal, selle arvelt laieneb mitmete teiste käänete kasutus. **Nimetava käände** esinemus ei erista A2- ja B1-taset, vähenedes seejärel B2- ja C1-tasemel nii nimi-, omadus- kui ka asesõnade lõikes. Keeleoskustasemega on tugevaimas negatiivses korrelatsioonis nimetava käände summaarne sagedus, mille hajuvus on tasemeti kõige väiksem. Nimisõnade puhul on nimetav kääne teksti tasemega seotud üsna nõrgalt ($\rho = -0{,}379$), põhjuseks suhteliselt suur varieerumine A2–B2-taseme kirjutistes ja keskmiste osakaalude väiksem erinemine võrreldes omadus- ja asesõnadega.

Avaram käändekasutus seostub tegusõnade leksikaalse mitmekesisusega, mis suureneb verbivarieeruvuse indeksitele tuginedes A2–C1-tasemel järjepidevalt (Allkivi-Metsoja 2021: 25–27). Nimetava käände ületarvitamist tingib tegusõna *olema* domineerimine, iseäranis A2-taseme tekstides. Kõigil tasemetel on *olema* laiendiks valdavalt nimetavaline öeldistäide (*sa oled väga professionaalne*).

Ülejäänud grammatiliste käänete osakaal kasvab keeleoskustaseme tõustes. **Omastav kääne** eristab B1–C1-taset, kuid omastavaliste asesõnade puhul suureneb üksnes B1- ja B2-taseme võrdluses, seostudes teksti tasemega tunduvalt nõrgemalt ($\rho = 0{,}278$) kui omastava käände sagedus nimi- ja omadussõnadel.

Järk-järgult laieneb omastavaliste täiendite kasutus tekstis, B2- ja C1-tasemel esineb varasemast rohkem mitut eestäiendit sisaldavaid nimisõnafraase (*selle firma töö*, *parema elukvaliteedi otsimine*). Lisaks sageneb omastava käände vormide kasutus tagasõna laiendina (*kaheksa tunni jooksul*, *oma pere piires*) ja eeskätt C1-tasemel täissihitisena (*unustad turvavöö kinnitada*).

Omastava käände sagedust võivad mõjutada sihitisekäände asendused: omastavalise täissihitise asendamine nimetavalisega (peamiselt A2–B1-tasemel, nt *ta tahab avada *teine kauplus ~ teise kaupluse*) või osasihitisega (kuni B2-tasemeni, nt *viskan sulle *päästerõngast ~ päästerõnga*, *sa saad mulle *nõuannet ~ nõuande anda*). Need kuuluvad õppijakeele sagedamate käändeasenduste hulka (vt Eslon 2011). Vilunud õppijate tekstides tuleb niisuguseid asendusi ette harvem.

Kui nimetava kasvav ja omastava kahanev kasutus ilmneb ka teksti sõnede suhtes arvutatud osakaalude põhjal (Allkivi-Metsoja 2021: 38 jj), siis **osastava käände** sagedus järjestikuseid keeleoskustasemeid oluliselt ei piiritle. Vaadeldes osastava vormide osakaalu kõigi käändsõnade suhtes, tuleb esile erinevus B1- ja B2-taseme kirjutiste vahel. Ka asesõnade kasutus osastavas sageneb B2-tasemel, nimisõnade puhul toimub muutus aga B1-tasemel. Kuna osastava käände kasutus eristab ühte järjestikust tasemepaari, siis on korrelatsioon keeleoskustasemega nõrk (asesõnadel $\rho = 0{,}396$; käändsõnadel summaarselt $\rho = 0{,}315$; nimisõnadel $\rho = 0{,}219$).

Omadussõnadel on muutus erisuunaline: osastava käände keskmine osakaal on A2-tasemel 11,7% (SH 20,7), B1-tasemel 7,5% (SH 13,2), B2-tasemel 7,1% (SH 13,2)

ja C1-tasemel 14,9% (SH 10,4). See suureneb küll C1-tasemel statistiliselt olulisel määral, kuid ei erine A2-tasemest.

Peamiselt A2- ja B1-tasemel tuleb ette nimetava käände kasutamist osastava asemel, eelkõige sihitise vormi valikul (*külastame *mingi muuseum ~ mingit muuseumi*, *õppisin prantsuse *keel ~ keelt*). Leidub ka sihitise ja eestäiendi ühildumise vigu, nt *mul on kaks *väike ~ väikest last*, *mul oli palju erinevaid *loomad ~ loomi*.

Kõrgemate tasemete suunas liikudes avarduvad osastava käände süntaktilised funktsioonid. Osasihitise kõrval (*ma igatsesin sind*, *kuulasime kontserti*) muutub B1–C1-tasemel üha sagedamaks osastava vormide kasutamine osaaluse funktsioonis peamiselt koos tegusõnaga *olema* (*pole kahtlustki*), kuid ka muude tegusõnadega (*midagi läheb katki*, *et seda ei juhtuks*). Osastav kääne esineb samuti hulgasõna laiendina (*kolm aastat*, *palju inimesi*), öeldistäitena (*ta oli pikka kasvu*) ja tervitusväljendites (*Head aega!*, *Kõike head!*), B1-tasemest alates lisaks ajamääruslikes kaassõnafraasides (*pärast tööpäeva*, *enne seda projekti*) ning B2- ja C1-tasemel väljendtegusõnades, nagu *aru saama*, *osa võtma*, *lugu pidama*, *huvi pakkuma/tundma*.

Nimetava, omastava ja osastava käände üldine sagedus läheneb järk-järgult eesti kirjakeele sagedusandmetele, olles C1-taseme tekstides sarnane TK andmetega. Ka nimisõnade puhul on grammatiliste käänete osakaalude jaotumine C1-tasemel kirjakeele lähedane. Omadussõnadel jääb nimetava käände osakaal siiski suuremaks ja omastava osakaal väiksemaks kui TK-s (osastava osakaal on sarnane: 13,6%). Asesõnadel on nimetava ja omastava käände kasutus TK-ga kõige sarnasem B2-tasemel, C1-tasemel kujuneb nimetav mõnevõrra harvemaks ja omastav sagedamaks (ilmselt omajatäiendite *oma*, *meie* jt ohtra kasutuse tõttu). Osastava käände kasutus on C1-tasemel endiselt veidi väiksem, kui kirjakeeles täheldatud.

3.4. Semantiliste käänete kasutus

Keeleoskustasemega on kõige tugevamalt seotud **saav kääne**, mille osakaal kasvab peamiselt B1–C1-tasemel. Kõigi käändsõnade lõikes ja nimisõnadel eristab see statistiliselt olulisel määral ka A2- ja B1-taset. Samal ajal on vahe väga väike ja seda ei tasu üldistada. A2-tasemel esineb saav kääne ühesainsa korra, ent ka B1-tasemel vaid 22 tekstis (nimisõnavorm 14, omadussõnavorm 9 ja asesõnavorm 2 tekstis). B2-tasemel leidub saavat käänet rohkem kui pooltes tekstides (69), omadus- ja asesõnadega siiski vähestes kirjutistes (vastavalt 21 ja 17). Oluline erinevus ilmneb alles C1-tasemel, kus saavat käänet esineb peaaegu kõigis tekstides (112), omadus- ja asesõnadega vastavalt kahes kolmandikus ja pooltes tekstides. Nii on käände üldine sagedus ja nimisõna saava käände kasutus teksti tasemega tugevamas korrelatsioonis.

Sagedamini tarvitatakse saavat käänet kokkuvõtte sõnastamisel (*kokkuvõtteks*, *lõpetuseks*); seisundimäärusena koos tegusõnadega *olema*, *muutuma*, *saama*, *jääma* ja *tegema* (*põhiliseks teemaks oli tervis*, *teenustasu muutus suuremaks*, *soovib edukaks saada*) ning ühendsidendis *selleks et*, harvem muudes otstarbemäärustes (*probleemide lahendamiseks*); määrustäiendina (*võimalused õnnetuste vältimiseks*) ja ajamäärusena (*üheks õhtuks*). Nimi- ja omadussõnadel on saava käände esinemus C1-tasemel sarnane TK-ga, asesõnadel suuremgi.

Seestütleva käände osakaal eristab keeleoskustasemeid selgeimalt asesõnade lõikes, suurenedes B1–C1-taseme võrdluses, ning kasvab ka omadussõnadel, kuigi ei erista järjestikuseid tasemeid, vaid C1-taset A2- ja B1-tasemest (A2 0,2%, SH 1,6; B1 0,6%, SH 2,9; B2 1,6%, SH 9,8; C1 3,3%, SH 4,5). Nimisõnade tarvitamine seestütlevas käändes sageneb A2-tasemelt (0,7%, SH 2,3) B1-tasemele liikudes (5,0%, SH 6,3), väheneb B2-tasemel (2,5%, SH 3,7) ja sageneb taas C1-tasemel (4,3%, SH 2,9).

Vastandlik tendents tuleneb B1-taseme eksamiülesandest, kus kirjeldatakse meediakanaleid, millest infot saadakse, ja teemasid, millest huvitutakse (*otsin infot internetist*; *loen ajakirju, mis on ajaloost või bioloogiast*). Selle eksami tekstides on nimisõna seestütleva vormide osakaal keskmiselt 11,0%, ülejäänud eksamikordi arvestades aga 3,0%. Erisuunaline muster avaldub seetõttu ka seestütleva käände sageduses tervikuna (A2 0,7%, SH 1,6; B1 3,1%, SH 3,8; B2 2,0%, SH 2,3; C1 4,2%, SH 2,2). Ometi on see tunnus keeleoskustasemega tugevamalt seotud kui nii mõnedki samas suunas muutuvad tunnused (vt tabel 2). Üldistatavaks võib pidada seestütleva käände suurenevat kasutust B1- ja C1-tasemel. Vilunud õppijate kirjutistes on seestütleva esinemus summaarselt ning nimi- ja omadussõnadel väga sarnane TK statistikaga (vastavalt 3,9%, 4,2% ja 3,4%), asesõnadel sellega võrreldes pisut suurem. Nimisõnu esineb seestütlevas käändes pea kõigis C1-taseme tekstides, asesõnu kolmveerandis ja omadussõnu pooltes tekstides.

Seestütleva funktsioonid laienevad järjepidevalt. A2-tasemel esinevad seestütlevas käändsõnad aja- ja kohamäärusena (*esmaspäevast reedeni*, *tuleme tagasi Riiast*), alates B1-tasemest sõltuvusmäärusena (*rääkisime elust*) ja määrustäiendina (*ajakirjad moest*), B2-tasemest hulgasõna laiendina (*enamik inimestest*), C1-tasemel lisaks ka eestäiendina (*metallist navigatsiooniseadme*) ja võrdlusalusena võrdlustarindis (*ID-kaardi kasutamisest turvalisemat*).

Seesütleva käände kasutus sageneb omadussõnadel C1-tasemel ning asesõnadel B2- ja C1-tasemel. B1- ja B2-taseme erinevus on siiski praktiliselt ebaoluline, sest enamikus B2-taseme tekstides (89) asesõnu seesütlevas käändes ei leidu. Seevastu väheneb B2-tasemel nimisõnade osakaal seesütlevas (A2 12,4%, SH 13,1; B1 10,7%, SH 6,7; B2 7,1%, SH 4,9; C1 7,1%, SH 3,6), kuid suure hajuvuse tõttu puudub tunnusel seos keeleoskustasemega. Osaliselt tuleneb varieerumine teksti teemast, seostudes kohamäärustega. Oluline erinevus tuleb esile A2-tasemel, kus nimisõna seesütleva käände vorme esineb ohtramalt reisikirjeldustes ja -kutsetes (*käisin muuseumis*, *jalutame Riias kesklinnas*), ning B2-tasemel, kus need on sagedaimad töökohta kirjeldavates tekstides (*mulle meeldib töötada kollektiivis*, *Atmosfäär meie asutuses on ülihea*).

Kokkuvõttes muutub seesütleva käände esinemus eri suundades, kahanedes mõnevõrra B2-tasemel ja kasvades uuesti C1-tasemel (A2 6,6%, SH 7,4; B1 5,7%, SH 3,6; B2 4,3%, SH 3,1; C1 5,7%, SH 3,2). See on suhteliselt sarnane TK andmetega (5,0%). Käände funktsioonid mitmekesistuvad eelkõige B2-tasemel: koha- ning ajamääruste (*alguses*, *tulevikus*) kõrval hakatakse selle vorme sagedamini tarvitama seisundimäärusena (*kursis*, *hädas*), lisanduvad sõltuvusmääruse (*saab mind selles aidata*) ja määrustäiendi funktsioon (*väikesed muudatused elustiilis*). C1-tasemel esinevad seesütlevas ka eestäiendid (*joobes autojuhid*), viisi- ja hulgamäärused (*kiiremas korras*, *suures osas*).

Nõrgalt on keeleoskustasemega seotud asesõna **kaasaütleva käände vormid**, mille suurem kasutus eristab C1-taset eelnevatest (ρ = 0,332). C1-taseme kirjutistes on kaasaütlev kääne asesõnade lõikes sagedam kui TK andmetes, esinedes eelkõige sõltuvusmäärusena tegusõnadega *nõus olema*, *jagama*, *tegelema*, *seotud olema* jm (*Ma olen sellega nõus*, *Tahaks jagada teiega oma arvamusi*).

Samal ajal väheneb kaasaütleva käände osakaal nimisõnadel, kuid seda ei saa lugeda kahaneva väärtusega tunnuseks, sest B2-tasemeni kaasaütleva sagedus vähehaaval suureneb ja C1-taseme tekstid ei erine oluliselt A2-taseme omadest (A2 3,7%, SH 6,5; B1 4,8%, SH 4,4; B2 6,0%, SH 3,8; C1 2,9%, SH 2,2). Analoogne erisuunaline muutus kajastub käände üldises osakaalus, mis ei erista C1-taset B1- ja A2-tasemest (A2 2,5%, SH 3,8; B1 3,0%, SH 2,5; B2 3,8%, SH 2,3%; C1 2,5%, SH 1,5). Korrelatsioon teksti tasemega puudub. TK statistikas on kaasaütleva sagedus võrdlemisi sarnane (summaarselt 2,8%, nimisõnadel 3,9%). Omadussõnu tarvitatakse selles käändes üliharva (osakaal C1-tasemel ja TK-s ühtviisi 0,2%).

Kaasaütlevat tarvitatakse juba A2-tasemel mitmes funktsioonis: kaasnemis-, vahendi- ja sõltuvusmäärusena (*mängisime lastega rannas*, *sõidame bussiga*, *ma tegelen reklaamiga*), määrustäiendina (*linna nimega Reykjavik*) ja kirja lõpetavas tervitusvormelis (*parimate soovidega*). B1-tasemel lisanduvad ja B2-tasemel sagenevad viisi- ja seisundimäärused (*jookseb rõõmuga minu juurde*, **asfaalt ~ asfalt on kahjustustega*) ning eestäiendid (*peaksite pakkuma tähtajaga lepingut*). C1-tasemel esineb kaasaütleva käände vorme tingimus- ja ajamääruse funktsioonis (*tiheda liiklusega on suurenenud liiklusõnnetused*, *oskus kujuneb aastatega*), ent varasemast vähem kasutatakse kaasnemismäärusi.

Sagedamini esinevate käänete hulka kuuluvad veel alale- ja alalütlev ning sisseütlev, mille sagedus ei seostu keeleoskustasemega, mistõttu nende käänete kasutusel siin pikemalt ei peatuta.

Läbivalt on marginaalne alaltütleva, rajava, oleva ja ilmaütleva käände esinemus. Nende osakaal jääb nii summaarselt kui ka eri sõnaliikidel tasemest olenemata alla 1%, mis on kooskõlas TK andmetega.

3.5. Omadussõna võrdevormide kasutus

Omadussõna võrdevormide kasutus erineb keeleoskustasemeti. Keskvõrde osakaal suureneb B1- ja C1-tasemel, eristamata samas B2-taset B1-tasemest. Selle arvelt väheneb algvõrde osakaal. Muutus on statistiliselt oluline A2- ja B1-taseme võrdluses.

Kui A2-taseme kirjutistes keskvõrret peaaegu ei esine, siis B1-tasemel on seda kasutatud kolmandikus, B2-tasemel ligi pooltes ja C1-tasemel valdavas osas tekstidest. Omadussõnade keskvõrde vorme tarvitatakse lisaks eestäiendi ja öeldistäite funktsioonile (*vanem tütar*, *kõigil oleks parem ja lõbusam*) samuti võrdlustarindites (*kergem kui teised tööd*), seisundimäärusena (*kasvavad suuremaks*) ja ülivõrde liitvormi osana (*kõige olulisem küsimus*), C1-tasemel ka sõltuvusmäärusena (*jäädakse rahule vähemaga*) ja täiendina hulgafraasis, kus hulgasõna on *üks*, nt *üks kasulikumaid ja huvitavamaid leiutisi*.

Ülivõrde vorme enamikus tekstides ei leidu (keskmine osakaal: A2 0,4%, SH 2,9; B1 0,3%, SH 2,2; B2 0,9%, SH 3,4; C1 0,5%, SH 1,7). Eelistatakse liitülivõrret ning

lihtülivõrde asemel kasutatakse ekslikult keskvõrret (*ta oli minu *parem ~ parim sõber*).

Omadussõnade algvõrde osakaal on alates B1-tasemest kirjakeelega sarnane, keskvõrre on C1-tasemel pisut sagedam ja ülivõrre harvem (osakaal TK-s 1,3%).

Eelnevast järeldub, et keeleoskustasemega seotud käändsõnakasutuse tunnused eristavad eelkõige ühte või kahte järjestikust tasemepaari – sagedamini on need B1–B2 ja B2–C1. Vähem on neid käändsõnatunnuseid, mis piiritlevad A2- ja B1-taset. Käändevormide funktsioonid laienevad enamasti igal järgneval tasemel, isegi kui vormi sagedus olulisel määral ei muutu.

4. Kokkuvõttev arutelu

Uurimuse eesmärk oli keeleõppijate eksamikirjutiste põhjal esile tuua käändsõnade vormikasutust iseloomustavad arvtunnused, mis võimaldavad A2–C1-taseme tekste üksteisest eristada. Selleks mõõdeti korrelatsioonanalüüsi abil tunnuste seost keeleoskustasemega ja kontrolliti erinevuste statistilist olulisust.

Keeleoskustasemega on tugevas korrelatsioonis käändsõnatunnused, mille väärtus kasvab või kahaneb püsivalt, eristades kõiki järjestikuseid tasemete paare: A2–B1, B1–B2 ja B2–C1. Läbivalt suureneb tekstis esinevate käändevormide arv (ka sõnaliigiti), sageneb mitmuse- ja väheneb ainsusevormide kasutus (v.a asesõnadel, mille kasutuses ilmneb muutus B1–C1-tasemel). Käänete arv kõigub tasemete piires vähe: erinevus keskmisest piirdub nii tervikuna kui ka eri sõnaliikide lõikes ühe-kahe käändega. Siiski ei saa selle tunnuse alusel tasemete vahele selgeid piire tõmmata. Ainsuse ja mitmuse osakaal sõltub kirjutise teemast ja varieerub tasemete piires märkimisväärselt (vt Allkivi-Metsoja 2021: 39), ehkki iga taseme keskväärtused erinevad suhteliselt suurelt. Keeleoskustaset hinnates on neid tunnuseid seega otstarbekas kombineerida muude käändsõnatunnustega, mis aitavad eristada kindlaid tasemepaare.

Kahte või ühte tasemepaari piiritlevate tunnuste hulka kuuluvad käändevormide ja omadussõna võrdlusastmete osakaalud (seos keeleoskustasemega on keskmine või nõrk). Käänetest osutusid tasemete eristamisel olulisimaks ühelt poolt nimetav, mille esinemus kahaneb B2- ja C1-tasemel ühtviisi nii nimi-, omadus- kui ka asesõnadel, ning teisalt saav ja omastav, mille kasutus laieneb samuti kõigis kolmes sõnaliigis. Erandlikult on teksti tasemega keskmise tugevusega korrelatsioonis seestütlev kääne, mille kasutus ei suurene läbivalt, vaid väheneb B1- ja B2-taseme võrdluses. Jättes kõrvale ühe B1-taseme eksami kirjutised, kus seestütlev esineb teemast tulenevalt märksa sagedamini kui ülejäänud tekstides, võib üldistada erinevuse A2- ja B1- ning B2- ja C1-taseme vahel. Kirjutamisülesande mõju seestütleva jt semantiliste käänete kasutusele tuleks edaspidi mahukama valimi põhjal lähemalt uurida.

Nimetava käände ja ainsuse esinemus on keeleoskustasemeid eristavate tunnustena esile tulnud ka saksa õppijakeele automaatanalüüsis – need muutelõputa ja suure sagedusega grammatilised vormid taanduvad taseme tõustes, samas sageneb muude käändevormide (genitiivi ja daativi) kasutus (Hancke 2013: 49–51; Szügyi jt 2019: 35). Käändekasutus avardub koos sõnavara mitmekesistumisega (vt Allkivi-Metsoja 2021): uued sõnad ja kasutuskontekstid tingivad

ka uute käändevormide tarvitamise ja käänete kasutamise mitmekesisemates funktsioonides.

Siinsed tulemused kinnitavad, et käändsõnade vormikasutust on kasulik vaadelda mitte ainult summaarselt, vaid ka sõnaliikide kaupa.

a) Summaarses ja eri sõnaliikide arvestuses muutub käändevormide osakaal erinevalt. Mitmekesistuv käändekasutus avaldub erinevatel tasemetel: saav kääne eristab üldiselt B1–C1-taset, kuid omadus- ja asesõnade puhul B2- ja C1-taset; omastav kääne eristab üldiselt B1–C1-taset, ent asesõnade osas vaid B1- ja B2-taset; osastav kääne eristab üldiselt B1–B2-taset, nimisõnadel aga A2- ja B1-taset; seestütlev eristab asesõnade puhul B1–C1-taset, omadussõnadel aga ei erista järjestikuseid tasemeid.
b) Muutus käänete kasutuses võib ilmneda ainult kindla sõnaliigi piires. Seesütlev kääne sageneb asesõnadel B2- ja C1-tasemel ning omadussõnadel C1-tasemel, kaasaütleva käände kasutus suureneb C1-tasemel vaid asesõnadel.

Võrdlus TK andmetega toob esile, et tase-tasemelt läheneb käändsõnavormide kasutus keeleõppijate tekstides eesti kirjakeelele. Huvitava suundumusena ilmneb mitmuse suurem sagedus C1-taseme tekstides, mille põhjus vajab eraldi uurimist. Üks võimalik seletus on kirjutamisülesande mõju (arutlemine ja üldistuste tegemine ühiskondlikel teemadel).

Teine suurem erinevus TK andmetest seostub omastava käände kasutusega. C1-taseme kirjutisi iseloomustab omastava väiksem osakaal nimi- ja iseäranis omadussõnadel, samas suurem osakaal asesõnade puhul. Omastava käände vormid esinevad valdavalt eestäiendina, sihitisena pruugitakse neid C1-tasemelgi harva. Nii tõuseb omastav kääne rohkem esile asesõnadel omajatäiendite (peamiselt *oma* ja *meie*) sageda kasutuse tõttu. Ka eesti keelt emakeelena kõnelejate keelelistes valikutes on täheldatud omastava käände kasutusala kitsenemist ja taandumist teiste sihitisekäänete kõrval (Ehala 2009, Eslon, Õim 2010).

Üksiktekstidele tugineva analüüsi tulemused ei anna ülevaadet vastava taseme õppijate kirjalikust keelekasutusest selle täies mitmekesisuses. Sellegipoolest saab siinse tekstivalimi põhjal teha järeldusi selle kohta, milliseid keelevahendeid ühe või teise taseme tekstiloomes aktiivsemalt kasutatakse ja mis funktsioonides need esinevad. Sellised andmed võimaldavad täpsustada keeleoskustasemete kirjutamise osaoskuse kirjeldusi.

Näiteks põgus võrdlus täiskasvanud õppija A1- ja A2-taseme grammatikapädevuse üldiste kirjeldustega (Kallas jt 2020) viitab, et osa madalamate keeleoskustasemetega seostatud grammatilisi vorme tuleb eesti keele õppija tekstides esile hoopis kõrgematel tasemetel. Nii on seestütleva käände kasutus lähtekoha väljendamiseks ära toodud juba A1-taseme kirjelduses, kuid seda leidub A2-tasemel vaid üksikutes tekstides ning alles B1-tasemest alates sisaldab seestütlevat suurem osa tekste. A2-taseme kirjeldus hõlmab saava käände tarvitamist otstarbe, seisundi ja tähtaja funktsioonis, kuid see käändevorm esineb veel B1-tasemel vähestes kirjutistes, tulles laiemalt kasutusse alles B2-tasemel.

Tekstide automaathindamise jaoks on vaja teada, millised muutevormid tegelikult eri tasemete tekstides avalduvad ja millise sagedusega. Siinne uurimus annab sellist teavet käändsõnavormide kasutuse kohta. Eksamitekstide analüüsi tulemusel selgitati välja rida olulisi, teksti taset iseloomustavaid käändsõnatunnuseid, mida

on kavas arvesse võtta keeleoskustaset ennustavate klassifitseerimismudelite koostamisel. Kui täpselt saab nende tunnuste alusel eksami- jm tekstide taset prognoosida, selgub edaspidi. Tõenäoliselt on tulemuslikum käändsõnatunnuste kombineerimine tegusõnakasutust ilmestavate tunnustega, samuti leksikaalsete ja süntaktiliste tunnustega.

Viidatud 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 four-grams’ – Lähivõrdlusi. Lähivertailuja, 26, 54–83. https://doi.org/10.5128/LV26.02

Allkivi-Metsoja, Kais 2021. Eesti keele A2–C1-taseme kirjalike tekstide võrdlev automaatanalüüs [‘Written Estonian at the levels A2–C1: Comparative automated analysis’]. – Lähivõrdlusi. Lähivertailuja, 31, 13–59. https://doi.org/10.5128/LV31.01

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; Ballier, Nicolas; Gaillat, Thomas; Lissón, Paula 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 (12.5.2022).

Delacre, Marie; Leys, Christophe; Mora, Youri L.; Lakens, Daniël 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

Ehala, Martin 2009. Keelekontakti mõju eesti sihitiskäänete kasutamisele [‘Linguistic contacts and the use of object cases in Estonian’]. – Keel ja Kirjandus, 3, 182–204.

Eslon, Pille 2008. Käändevormide kasutussageduse võrdlus eesti õppijakeeles ja kirjakeeles [‘Comparison of case form frequency in learner and standard Estonian’]. – Pille Eslon (Toim.), Õppijakeele analüüs: võimalused, probleemid, vajadused. Tallinna Ülikooli eesti filoloogia osakonna toimetised 10. Tallinn: TLÜ Kirjastus, 31–66.

Eslon, Pille 2011. Millest räägivad eesti õppijakeele käändeasendused? [‘Implications of the Estonian learner language case replacements’]. – Lähivõrdlusi. Lähivertailuja, 21, 45–64. http://doi.org/10.5128/LV21.02

Eslon, Pille; Matsak, Erika 2009. Eesti keele kasutusvariandid: korpustest tulenev käändevormide võrdlev analüüs [‘Corpus-driven comparative analysis of variants of Estonian’]. – Eesti Rakenduslingvistika Ühingu aastaraamat, 5, 79–110. http://doi.org/10.5128/ERYa5.06

Eslon, Pille; Õim, Katre 2010. Objektikäänete kasutamisest sageduse ja markeerituse seisukohalt [‘About Estonian object cases from the perspective of markedness and frequency’]. – ESUKA/JEFUL, 1 (2), 69–89. https://doi.org/10.12697/jeful.2010.1.2.05

Eslon, Pille; Kaivapalu, Annekatrin 2020. Teel sihtkeelepärase keelekasutuse poole: vene- ja soomekeelsete eesti keele õppijate kirjaliku keelekasutuse dünaamika A2- ja B1-tasemel [‘Towards target-like language use: Russian and Finnish learners’ dynamics of written Estonian on A2- and B1-level’]. – Lähivõrdlusi. Lähivertailuja, 30, 57–88. http://doi.org/10.5128/LV30.01

Guilford, Joy Paul 1973. Fundamental Statistics in Psychology and Education. New York, NY: McGraw-Hill.

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

Hawkins, John A.; Buttery, Paula 2010. Criterial Features in Learner Corpora: Theory and Illustrations. – English Profile Journal, 1 (5), E5. https://doi.org/10.1017/S2041536210000103

Kallas, Jelena; Koppel, Kristina; Üksik, Tiiu 2020. Eesti keele kui teise keele õpetaja tööriistad. Eesti Keele Instituut. https://doi.org/10.15155/3-00-0000-0000-0000-08357L

Kerz, Elma; Wiechmann, Daniel; Qiao, Yu; Tseng, Emma; Ströbel, Marcus 2021. Automated classification of written proficiency levels on the CEFR-scale through complexity contours and RNNs. – Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 199–209.

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.

McKinney, Wes 2010. Data structures for statistical computing in Python. – Proceedings of the 9th Python in Science Conference, 56–61. https://doi.org/10.25080/Majora-92bf1922-00a

Meurers, Detmar 2019. Natural Language Processing and Language Learning. – Carol A. Chapelle (Toim.), Concise Encyclopedia of Applied Linguistics. Wiley-Blackwell, 817–831.

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

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

Pool, Raili 2007. Eesti keele teise keelena omandamise seaduspärasusi täis- ja osasihitise näitel ['The Acquisition of Total and Partial Objects by Learners of Estonian as a Second Language']. Dissertationes philologiae estonicae Universitatis Tartuensis 19. Tartu: TÜ Kirjastus.

Qi, Peng; Zhang, Yuhao; Zhang, Yuhui; Bolton, Jason; Manning, Christopher D. 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages. – Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 101–108. https://doi.org/10.18653/v1/2020.acl-demos.14

Rysová, Katerina; Rysová, Magdaléna; Novák, Michal; Mírovský, Jirí; Hajičová, Eva 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

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

Tack, Anaïs; Francois, Thomas; Roekhaut, Sophie; Fairon, Cédrick 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. https://doi.org/10.18653/v1/W17-5018

Üksik, Tiiu; Kallas, Jelena; Koppel, Kristina; Tsepelina, Katrin; Pool, Raili 2021. Estonian as a second language teacher's tools. – Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, 130–134.

Vajjala, Sowmya 2018. Automated assessment of non-native learner essays: Investigating the role of linguistic features. – International Journal of Artificial Intelligence in Education, 28, 79–105. https://doi.org/10.1007/s40593-017-0142-3

Vajjala, Sowmya; Lõo, Kaidi 2014. Automatic CEFR level prediction for Estonian learner text. – Proceedings of the 3rd Workshop on NLP for Computer-assisted Language Learning. NEALT Proceedings Series 22, 113–127.

Volodina, Elena; Pilán, Ildikó; Enström, Ingegerd; Llozhi, Lorena; Lundkvist, Peter; Sundberg, Gunlög; Sandell, Monica 2016. SweLL on the rise: Swedish Learner Language corpus for European Reference Level studies. – Proceedings of the 10th International Conference on Language Resources and Evaluation, 206–212.

Voolaid, Katrin 2018. Vene ja soome lähtekeelega õppijate eesti keele kasutus- mustrid (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

Yannakoudakis, Helen; Andersen, Øistein E.; Geranpayeh, Ardeshir; Briscoe, Ted; Nicholls, Diane 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

Lisa 1. Käändsõnade vormikasutuse erinevused Welchi F-statistiku alusel

Kasutatud on Bonferroni meetodil korrigeeritud olulisusnivood 0,001. Rühmadevaheline vabadusastmete arv k – 1 = 3, kuna võrreldavate rühmade arv k = 4.

* Tunnus eristab kõrvuti asetsevaid tasemeid, kuid korrelatsioon keeleoskustasemega puudub.
** Tunnus ei erista kõrvuti asetsevaid tasemeid.

Tunnuse liik	Tunnus	Welchi F	Rühmadesisene vabadusastmete arv	p-väärtus
Summaarsed tunnused	Käändevormide arv	249,3	264,1	< 0,001
	Nimetav kääne	141,4	261,8	< 0,001
	Omastav kääne	103,0	261,1	< 0,001
	Osastav kääne	16,7	260,1	< 0,001
	Sisseütlev kääne*	6,1	256,4	0,001
	Seesütlev kääne*	6,7	258,3	< 0,001
	Seestütlev kääne	71,8	255,2	< 0,001
	Alaleütlev kääne*	7,6	256,6	< 0,001
	Alalütlev kääne*	5,3	258,6	0,001
	Alaltütlev kääne*	7,5	257,3	< 0,001
	Saav kääne	88,3	245,3	< 0,001
	Rajav kääne*	7,1	247,8	< 0,001
	Ilmaütlev kääne**	7,1	261,4	< 0,001
	Kaasaütlev kääne*	9,1	253,4	< 0,001
	Ainsus / Mitmus	230,9	261,3	< 0,001
Nimisõnatunnused	Käändevormide arv	266,3	264,3	< 0,001
	Nimetav kääne	46,5	251,8	< 0,001
	Omastav kääne	95,3	262,2	< 0,001
	Osastav kääne	10,0	254,1	< 0,001
	Sisseütlev kääne*	8,6	246,2	< 0,001
	Seesütlev kääne*	14,8	249,8	< 0,001
	Seestütlev kääne	45,7	254,4	< 0,001
	Alaleütlev kääne	3,4	258,8	0,019
	Alalütlev kääne	20,0	256,5	< 0,001
	Alaltütlev kääne*	7,9	247,0	< 0,001
	Rajav kääne*	8,7	227,7	< 0,001
	Ilmaütlev kääne	4,4	263,3	0,005
	Kaasaütlev kääne*	22,3	247,2	< 0,001
	Ainsus / Mitmus	280,8	256,4	< 0,001

Tunnuse liik	Tunnus	Welchi F	Rühmadesisene vabadusastmete arv	p-väärtus
Omadussõnatunnused	Käändevormide arv	237,1	259,4	< 0,001
	Nimetav kääne	71,3	249,4	< 0,001
	Omastav kääne	33,2	252,1	< 0,001
	Osastav kääne	13,1	248,2	< 0,001
	Seesütlev kääne*	19,2	233,5	< 0,001
	Seestütlev kääne**	17,0	236,6	< 0,001
	Alalütlev kääne	1,9	234,1	0,129
	Saav kääne	37,4	228,5	< 0,001
	Kaasaütlev kääne	0,2	255,8	0,903
	Ainsus / Mitmus	79,4	251,1	< 0,001
	Algvõrre	19,6	255,4	< 0,001
	Keskvõrre	22,6	255,6	< 0,001
	Ülivõrre	0,7	246,8	0,530
Asesõnatunnused	Käändevormide arv	242,8	263,1	< 0,001
	Nimetav kääne	86,0	258,0	< 0,001
	Omastav kääne	30,9	257,1	< 0,001
	Osastav kääne	27,4	259,9	< 0,001
	Seesütlev kääne	30,2	246,3	< 0,001
	Seestütlev kääne	42,8	257,2	< 0,001
	Alaleütlev kääne	11,6	256,2	< 0,001
	Alalütlev kääne	11,7	254,8	< 0,001
	Kaasaütlev kääne	6,9	261,6	< 0,001
	Ainsus / Mitmus	62,3	260,0	< 0,001

,

USE OF NOMINALS IN ESTONIAN A2–C1-LEVEL EXAM WRITINGS

Kais Allkivi-Metsoja
Tallinn University

In this study, natural language processing (NLP) is used to analyse nominal inflection in Estonian proficiency examination writings representing the CEFR levels A2–C1. The aim is to define the nominal features that distinguish learner language production at each proficiency level. For this purpose, the frequency and variation of inflectional forms are measured in two ways: a) for the nominal parts of speech (PoSs) in total, i.e., considering the use of nouns, pronouns, adjectives and numerals; b) for nouns, pronouns and adjectives individually (numerals were discarded due to low frequency).

The analysed corpus contains 480 texts, 120 for each level. Nominal features based on the grammatical categories of number, case and degree of comparison are extracted from the morphologically tagged and manually corrected output of the Stanza NLP toolkit. Relevant features are selected according to the following criteria: they correlate with the proficiency level, their values change monotonically, and there are statistically significant differences between (some) adjacent levels.

A2–C1-level texts are consistently distinguished by the number of cases used in the text as well as the ratio of singular and plural forms. The changes in the frequency of nominal inflectional forms mainly occur from level B1 to C1. The use of translative, nominative and genitive case are more strongly related to the text level, while partitive, inessive, elative and comitative case and comparative adjectives also differentiate some levels.

Furthermore, the study indicates that it is beneficial to observe inflection-based features separately for each PoS when analysing L2 development. Firstly, the PoS-specific frequencies of some grammatical categories increase at different stages of proficiency. Secondly, changes may emerge for certain PoSs only.

The identified criterial features could be used for automated assessment of Estonian L2 writings alongside lexical, syntactic and other linguistic features. The results can also help to specify the CEFR level descriptions for Estonian.

Keywords: natural language processing, morphology, CEFR levels, written learner language, Estonian

Kais Allkivi-Metsoja on Tallinna Ülikooli infoühiskonna tehnoloogiate doktorant, keskendub keeleoskustasemete statistilisele mudeldamisele.
Narva mnt 25, 10120 Tallinn, Estonia
kais@tlu.ee