TARTU ÜLIKOOL
FILOSOOFIATEADUSKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT
Eesti keele osakond

Kadri Sõrmus

Emakeeleõppija korpus

Statistiline analüüs ja veamärgendussüsteem

Magistritöö

Juhendajad Kadri Muischnek PhD
dots Maia Rõigas

TARTU 2008

Sisukord

Sissejuhatus **5**
1. Keelekorpus **8**
1.1. Õppijakeele korpus 9
1.2. Korpuse representatiivsus 13
2. Tekstivaliku põhimõtted **15**
2.1. Teksti karakteristikud 15
2.2. Autorist lähtuv tekstivalik 18
2.3. Situatsioonist lähtuv tekstivalik 18
2.4. Kogumisviisist lähtuv tekstivalik 19
2.5. Metaandmed 20
3. Korpus keeleõppes **22**
3.1. Korpuspõhised õppematerjalid 23
3.2. Õppijakeelt sisaldavad korpused 26
4. Korpuspõhised uuringud **33**
4.1. Märgendamata korpus 36
4.1.1. Sagedusloend 36
4.1.2. Konkordants 40
4.1.3. Kollokatsioon 42
4.2. Korpuse märgendamine 43
5. Õppijatekstide traditsiooniline uurimismeetod – veaanalüüs **49**
5.1. Viga 49
5.2. Veaanalüüsi etapid 52
5.2.1. Tekstide kogumine, parandamine ja veaklassidesse liigitamine 53
5.2.2. Märgendite süsteemi väljatöötamine ja korpuse märgendamine 57
5.2.3. Korpuse analüüs 60
6. Eesti kooliõpilaste kirjalik tekstikorpus **63**
6.1. Emakeeleõppijate tekstikorpus 63
6.1.1. Korpuse tekstid 63
6.1.2. Metaandmed 65

6.2. Emakeeleõppija korpuse statistiline analüüs 67
6.2.1. Üldandmed 67
6.2.2. Lausepikkus 67
6.2.3. Sagedusanalüüs 69
6.2.3.1. Õppijatekstides leiduvad sagedamad kollokatsioonid 69
6.2.3.2. Õppijatekstide sagedamad sõnavormid 70
6.2.3.3. Kahekümne sagedasema lemma analüüs 72
6.2.3.4. Asesõnad 73
6.2.3.5. Sidesõnad 73
6.2.3.6. Verbid 73
6.2.3.7. Nimisõnad 73
6.2.3.8. Hulgasõnad 73
6.2.3.9. *ei* ja *ka* 74
6.2.3.10. Sagedusanalüüs sõnaliigiti 75
6.3. Emakeeleõppija korpuse veakategooriad ja märgendussüsteem 84
6.3.1. Vigade jagamine üldkategooriatesse 84
6.3.2. Veakategooriad 85
6.3.3. Viga ja veapiir emakeeleõppija korpuses 87
6.3.4. Vigade märgendamine 88
6.3.4.1. Märgendid 88
6.3.4.2. Märgendamine 92
6.3.4.3. Veatüübid 96
6.3.5. Märgendatud vigade parandamine 128
6.4. Korpuse edasine arendustöö 132
Kokkuvõte 134
Corpus of Native Estonian Learners 136
Kasutatud materjalid 138
Lisad 145
Lisa 1. Näide korpuse kasutamist emakeeletunnis 145
Lisa 2. Näide korpuse kasutamisest emakeeletunnis 146
Lisa 3. Näide korpuse kasutamisest sõnavara õpetamisel 147
Lisa 4. Ülekasutatud sõnad õppijakorpuses 148

Lisa 5. Kõrgtasemel keeleõppijate vealiigid 149
Lisa 6. Gümnaasiumiõpilaste sagedamate keelevigade liigitus 150
Lisa 7. Sõnastusvead Leo Villandi järgi (1966)..... 151

Sissejuhatus

Kirjakeeleoskuse arengu suunamine on seotud keeleõppega üldhariduskoolis, kust saavad alguse eesti keele kasutamise korrektsus ja hoiakud keele suhtes. Eesti üldkirjakeele oskus ei ole üldhariduskooli lõpetanutel piisav. (Metslang 2004: 202) Õppimistulemuste parandamiseks on esmalt vaja õppija keelt analüüsida, leida enim probleeme tekitavad valdkonnad, et teadliku veaanalüüsi tulemustest lähtuvalt pöörata probleemsetele keelevaldkondadele suuremat tähelepanu nii ainekavade kui ka õppematerjalide koostamisel.

Keele omandamise uuringud hõlmavad kolme valdkonda:

> 1) väikelaste esimese keele (või esimese ja teise keele) omandamise uurimine;
> 2) teise keele või võõrkeele omandamise uurimine nii lastel kui täiskasvanutel;
> 3) emakeeleõppija keelekasutuse aspektide uurimine (näiteks lugemis- ja kirjutamisoskus erineva vanuseastme ja/või keeleoskuse tasemega laste keelekasutuses). (Biber *et al.* 172–173)

Õppijakeelt on võimalik uurida eksperimentaalselt ja vaatluslikult. Eksperimentaalne uurimine põhineb erinevatel testidel, mille abil selgitatakse välja õppija keeleoskuse tase uurimise hetkel. Vaatlusliku lähenemise korral fikseeritakse ühe või mitme informandi autentne keelekasutus ning seejärel analüüsitakse vaatluse tulemusi. Mahukas andmekogu muudab vaatlusliku analüüsi ulatuslikumaks.

Eestis on suurteks andmekogudeks näiteks Eesti Keele Instituudi kirjakeele arhiiv, murde- ja heliarhiiv, Tartu Ülikooli murdearhiiv jt. Õppijakeele uurimise vaatepunktist on omalaadne tekstikogu riikliku eksami- ja kvalifikatsioonikeskuse kirjandite arhiiv, mis sisaldab õpilaste kirjandeid alates aastast 1997, mil gümnaasiumi lõpukirjandite kirjutamine muutus riiklikult hinnatavaks eksamiks.

Suured andmekogud teevad huvipakkuvate tekstide kättesaamise hõlpsamaks, kuid ei võimalda analüüsi mingil moel automaatsemaks muuta. Keelekorpuste kasutuselevõtuga on vaatlusel põhinev analüüs saavutanud uue taseme – tagasihoidlike

andmehulkade kasutamise asemel baseeruvad tänapäevased lingvistilised uuringud suurtel elektroonilistel keelekogudel ehk korpustel. Eesti keelt säilitatakse elektroonilisel kujul TÜ eesti kirjakeele tekstikorpuses, vana kirjakeele tekstikorpuses ja suulise keele tekstikorpuses ning TÜ ja Eesti Keele Instituudi murdekorpuses. Kõik need korpused sisaldavad täiskasvanud keelekasutaja keelt.

Laste emakeelekasutuse fikseerimisele ei ole Eestis seni erilist tähelepanu pööratud. Praegu on Eesti väikelaste keele omandamise uurimiseks võimalik kasutada rahvusvaheliselt tuntud lastekeele korpust CHILDES (*Child Langue Data Exchange System*), mis sisaldab kümne eesti lapse osaliselt transkribeeritud ja osaliselt (uurija huvidest lähtuvalt) kodeeritud keelt. Samasse andmebaasi on kavas lisada ka kolmikute korpus. (Argus 2007: 66–67; Vija 2007: 374)

Eesti keelt teise keelena või võõrkeelena omandavate noorte terviktekste sisaldab Tallinna ülikooli võõrkeeleõppijate korpus ehk vahekeelekorpus. Praegu on huvilistel ligipääs keeleõppijate terviktekstidele ning teksti autori andmetele aadressil *evkk.tlu.ee*. P. Esloni juhtimisel tegeldakse eesti keele kui teise keele või võõrkeele õppija vigade analüüsimisega. Vahekeele korpus sisaldas 2007. aasta septembris 608 000 sõnet, millest märgendatud korpuse osakaal oli 217 000 sõnet. (Eslon 2007: 101)

Tartu Ülikooli juures kogutakse eesti keelt võõrkeelena õppivate tudengite keelevigu ning rajatakse vigade korpust. Arvutisse sisestatud vigaseid õppijakeele lauseid ning nende parandusi sisaldava korpuse maht on 2008. aasta jaanuarikuu seisuga 128 000 sõna.

Eesti emakeelega koolilapse keelekasutuse uurimiseks sobivat keelekorpust seni loodud ei ole. Ka tänapäeva eesti keele uurimusi ei ole korpuspõhiselt "keeleõppe eesmärgil üldistatud ega keeleõppe jaoks olulist lingvistilist teavet välja toodud" (Eslon 2007: 93).

Eesti keele osakonnas on õppijate keelekasutuse kohta kirjutanud bakalaureusetöö Taimi Vill 2001 „Viga eesti keeles eesti kooli õpilase hinnangul", Helen Treial 2004 „Morfoloogia- ja sõnamoodustusvead gümnaasiumi ning põhikoolilõpetajate riigieksamikirjandites", Elise Jüssi „Laadivahelduse taandumine eesti kirjakeeles õpilaste keelekasutuse näitel", Anni Jürine 2007 „Eesti kaassõna norminguvastasest

kokkukirjutamisest grammatiseerumisteooria taustal“. Korpuste kohta on K. Muischneck 1998. aastal kirjutanud magistritöö „Korpused ja nende kasutamine“.

Minu magistritöö eesmärgiks on luua eesti keelt emakeelena õppivate koolilaste kirjalike tekstide korpus (emakeeleõppija korpus), viia läbi selle esmane statistiline analüüs, keskenduda veaanalüüsiks sobiva märgendamise süsteemi loomisele ning teha ettepanekuid õppijakorpuse edasiseks arendustegevuseks. Veamärgendussüsteemi põhimõtted on lähtealuseks ka TÜ mitte-eesti emakeelega õppijate veakorpuse märgendamisele.

Eesmärgi saavutamiseks uurin õppijakorpuse loomise põhimõtted ja kogun õppijate tekste. Statistilise uuringu tegemiseks kasutan korpusanalüüsi meetodeid. Õppijate tekstides esinevate vigade märgendamise süsteemi loomiseks parandan 10 000 sõna õppijate tekste ning parandustest lähtuvalt töötan välja vigade liigitamise põhialused. Käesoleva töö eesmärk ei ole õpilastekstides esinenud vigu statistiliselt analüüsida.

Magistritöö esimene peatükk annab ülevaate korpuste liikidest, olemusest ja koostamisega seonduvast. Teine peatükk tutvustab tekstide valimise põhimõtteid õppijakorpuste loomisel. Kolmandas peatükis käsitletakse korpuste kasutamist keeleõppes. Neljas peatükk kirjeldab korpuspõhiseid uurimisvõimalusi. Viies peatükk tutvustab veaanalüüsi kui levinud lähenemisviisi õppijatekstide analüüsimiseks. Kuues peatükk sisaldab loodud emakeeleõppija korpuse kirjeldust, korpuse statistilist analüüsi ja õppijate vigade märgendamise süsteemi. Töö lisadest leiab näitlikku materjali.

Töö kirjutamisel olen lähtunud peamiselt Lynne Bowkeri, Susan Hunstoni, Silviane Grangeri, Douglas Biberi, Carl Jamesi ning Rod Ellise seisukohtadest. Nimetatud autorite tööd käsitlevad küll keeleõppijat peamiselt võõrkeele või teise keele õppijana, kuid samad põhimõtted kehtivad ka õppijate kohta üldiselt.

1. Keelekorpus

Keeleteaduses mõeldakse korpuse all tavaliselt kas kirjaliku või suulise keeleainese kogumit, mida on võimalik kasutada lingvistilise uurimistöö alusmaterjalina. Korpust võib nimetada ka andmete kogumise ja sorteerimise viisiks (Hunston 2002: 2; 20).

John Sinclairi (1995: 17) arvates peab korpuseks nimetatav keelekogu koosnema vaid kindlate kriteeriumite alusel kogutud tekstidest. Vabama tõlgenduse järgi võib korpuseks nimetada igasugust elektroonilisel kujul esinevat tekstikogu, mis on koostatud selleks, et kasutaja võiks talletatud tekstide hulgast vastavalt oma vajadustele eesmärgipäraseid valikuid teha. Nn kolmanda põlvkonna korpused on pigem elektroonilised tekstiarhiivid kui korpused, kuna nende koostamine on olnud juhuslik: kogutud on seda, mida on olnud võimalik elektroonilisel kujul kätte saada (Muischnek 1998: 8–12).

Elektrooniliste keelekorpuste loomise vastu on maailmas huvi tuntud umbes eelmise sajandi 60. aastatest alates, kui arvutite mälumaht hakkas võimaldama suurte tekstihulkade säilitamist. Alates selle aastatuhande algusest on Internetipõhine korpus ja selle analüüsivahendid muutunud järjest kättesaadavamaks ning korpuspõhised uuringud keeleteaduses täiesti tavaliseks (Biber *et al.* 2006: ix).

Korpust iseloomustatakse märksõnadega *autentne, elektrooniline, suur* ja *spetsiifiline* (Bowker, Pearson 2002: 9).

Autentsus (ingl *authentic*) iseloomustab korpustesse talletatud keelt: igapäevastest situatsioonidest kogutud tekstid, mille laused ei ole keelenähtuse demonstreerimiseks konstrueeritud. (Bowker, Pearson 2002: 9) ***Elektroonilisus*** (ingl *electronic*) tähistab seda, et valitud tekstikogu on arvutisse sisestatud ja arvutiprogrammide abil töödeldav: tekste on kergem kontrollida, hõlpsam analüüsida ja töötlemata andmetele kas automaatselt või pool-automaatselt vajalikku infot lisada (Granger 2007: 55). Märksõna ***suur*** (ingl *large*) tähistab korpuse mahtu. Enamasti määratakse korpuse suurust sõnade arvu järgi. Märksõna ***spetsiifiline*** (ingl *specific criteria*) lähtub keele variatiivsusest. Korpuse spetsiifilisus sõltub nii tekstide autorist, kirjutamise situatsioonist ja ülesandest kui ka teksti enda karakteristikutest (Tono 2002: 1). Üheks spetsiifiliseks korpuste liigiks ehk erikorpuseks on õppijakeele korpused.

Korpuste kasutamise eelistena on rõhutatud empiirilist, selgete printsiipide põhjal valitud materjalikogu, analüüsi kiirust, analüüsitava materjali mahukust, mis ühtlustab võimaliku valimivea ja muudab analüüsi objektiivsemaks (Biber *et al.* 2006: 3–9) ning võrreldavate kirjelduste loomise ja tekkinud hüpoteese kontrollimise lihtsamaks (Bowker, Pearson 2002: 21). Uurimisküsimustele vastuste saamiseks on keelekasutuse näiteid võimalik interpreteerida ja analüüsida varasemaga võrreldes teistsugusel viisil, kuid korpus ise vastuseid ei anna (Hunston 2002: 22–23). Korpuspõhine uurimine täiendab seniseid traditsioonilisi uurimismeetodeid, aga ei saa olla ainuke viis keeleuuringute tegemiseks.

Korpusanalüüsi meetodit kasutavad enim leksikograafid (sõnaraamatute, sõnastike koostamiseks), keeleõppijad ja õppematerjalide koostajad (autentse keelekasutusega tutvumiseks), sotsiolingvistid (keele varieerumise uurimiseks), keeleajaloolased (keele ajaloolise arengu uurimiseks), lingvistid (korpusepõhisteks analüüsideks), arvutilingvistid (tõlkesüsteemid jm loomuliku keele analüüs) (vt ka Bowker, Pearson 2002: 11; 21).

1.1. Õppijakeele korpus

Õppijakeele korpuseks (ka õppijakorpuseks) nimetatakse tekstide kogu (näiteks kontrolltööd, esseed, kirjandid, seletuskirjad vm), mille autoriteks on keeleõppijad. Enamasti käsitletakse õppijakorpust kui teise keele õppija (võõrkeeleõppija) tekstikorpust või spetsiifilist õppijakeele (e vahekeele) korpust, millele emakeelse õppija tekstid on lisatud vaid võrdlusalusena (Bowker, Pearson 2002: 13; Lemnitzer, Zinsmeister 2007: 115–126; Keeffe *et al*. 2007: 284–296, Tono 2002: 2). Käesolevas töös mõeldakse õppijakorpuste all emakeelt õppivate laste või noorte kirjutatud tekstidest koosnevat kirjaliku keele korpust. Täiskasvanud keeleõppijate, teise keele või võõrkeeleõppijate korpuste kohta käivat teavet on käsitletud õppija kohta üldiselt.

Korpuste kasutamise peamiseks eesmärgiks on välja selgitada, mis on keeles tüüpiline (Sinclair 1995: 17). Õppijakorpuste koostamise eesmärgiks on peegeldada õppijate keelele iseloomulikke jooni: tekstide ülesehitust, sõna- ning vormikasutust,

keele õppimise protsessi ja seaduspärasusi jm. Saadud teavet võib analüüsida nii teaduslikel kui ka keeleõpetuslikel eesmärkidel.

Õppijakeelt sisaldavaid ingliskeelseid korpuseid on hakatud koostama alates 1990. aastatest, seega ligi 30 aastat pärast üldkorpuste loomist (Nesselhauf 2005: 40). Teistel keeltel põhinevaid õppijakorpusi on hakatud koostama veelgi hiljem.

Laste keelekasutuse kogumine erineb täiskasvanute tekstide kogumisest. Materjali kogumist võivad takistada nii sobivate arvutisse sisestatud tekstide vähesus, tekstide kättesaadavus kui ka juriidilised probleemid.

I kooliastmel õppijate tekstikogu koostamine eeldab suurt sisestamistööd, kusjuures tekstid tuleb trükkida koos õppija tehtud vigadega ning ilma, et sisestaja omalt poolt vigu juurde lisaks. Algklassiõpilaste kirjutised on ka väga lühikesed, mistõttu peab neid vähegi arvestatava mahuga allkorpuse loomiseks koguma paljudelt õppijatelt.

Õppijakeele korpuse *Corpus of elementary student speech and writing* loonud Douglas Biber tunnistab (2006: 176), et emakeelsete õppijate tekste oli korpusesse väga keeruline hankida. Sobivaid ingliskeelseid tekste otsiti avalikest korpustest, millest kõige paremaks ja mahukamaks osutus CHILDESi keelekorpus. Childes sisaldab lisaks väikelaste keelele ka kooliealiste laste ja noorte suulist keelt. Õppijate kirjalikke tekste hangiti Grabe ja Reppeni loodud Arizona algkooliõpilaste korpusest *Arizona Corpus of elementary Student writing.* (Biber *et al*. 2006: 176–177)

Hõlpsamini kättesaadavad on gümnaasiumiealiste laste tekstid, näiteks need, mida õpilased saadavad kirjandi- või esseevõistlustele. Võistluskirjandid esindavad vaid selliste õppijate keelekasutust, keda võib tinglikult iseloomustada üldnimetusega *edasipüüdlikud* ning seetõttu ei anna ainult võistlustele saadetud töödest koostatud kogu objektiivset ettekujutust õppijate keelekasutuse kohta üldiselt.

Konkreetse uuringu jaoks sobivaim korpuse suurus sõltub uurija eesmärgist. Kui uurimise all on keeles harva esinev sõna, on sõnakasutuse eksponeerimiseks vaja oluliselt mahukamat korpust kui grammatiliselt tüüpilise nähtuse või sageliesineva sõna analüüsimiseks. (Bowker, Pearson 2002: 48. Biber *et al.* 2006: 269–274) Aktsepteeritava ning üldistava lingvistilise analüüsi tegemiseks peab korpus olema küllalt suur, et sisaldada piisaval hulgal keelenäiteid.

Õppijakorpuse ideaalset suurust ei ole võimalik määratleda. J. Sinclair (1995: 19) on väitnud, et kuna korpus sisaldab alati selliseid sõnu, mis esinevad seal vaid üks kord, ei ole ükski korpus liiga suur ning keele uurimiseks sobiv korpus peaks seetõttu sisaldama mitmeid miljoneid sõnu (Sinclair 1995: 19). Ka tavaline üldkeeles kasutatav sõna võib korpuses olla harva esinev ning vastupidi: kuna sagedus sõltub alliktekstidest, ei pruugi sagedus mingis tekstis või tekstiklassis garanteerida sõna tavalisust üldkeeles (Kaalep, Muischnek 2002: 9). Näiteks sõna *´crime´* (kuritegu), mida peetakse inglise keeles üsna tavaliseks, esines 1 miljoni sõna suuruses *Longman-Lancasteri* balansseeritud korpuses vaid 20 korda (Tribble 2008: 11).

Keeleõppijate korpuste seas peetakse 200 000-sõnalist korpust juba suureks, sest tavaliselt tuginetakse õppijakeele uuringutel palju väiksematele mahtudele. Üldises korpuste maailmas, kus mitmeid miljoneid sõnu sisaldavad megakorpused on üsna tavalised, on veerandmiljoniline korpus muidugi väga väike. (Granger 2003: 465) Mitmete allkorpuste üheks mahukaks korpuseks liitmisega on loodud ka miljoneid sõnu sisaldavaid õppijakorpuseid. Valdavalt sisaldavad need inglise keele õppijate tekste (Pravec 2002: 90) (vt ka tabel 5).

Korpuse suurusega on seotud ka selles sisalduvate tekstide arv. Kui koostada korpus näiteks 20 erineva autori tekstidest kogumahuga 20 000 sõna, siis pole see samaväärne, kui sisestada kaks sama autori teksti samuti kogumahuga 20 000 sõna. Erinevate autorite tekstivalimi kasutamine võimaldab uurida keelt üldisemalt. (Bowker, Pearson 2002: 49) Ühe autori tekstide põhjal saab analüüsida konkreetse autori keelekasutust (Davies 2007 :1).

Korpuse koostaja võib otsustada, et korpusesse pääsevad vaid need tekstid, mille sõnade arv ületab teatud piiri või jääb allapoole teatud piiri. Näiteks ingliskeelsesse õppijakeele korpusesse ICLE kogutakse tekste, mille pikkuseks on 500–1000 sõna, kuid jaapanlaste korpusse nimega JEFLL kogutakse õpilastöid pikkusega 20-150 sõna (Pravec 2002: 93–94). Algklassiõpilaste, kuid ka põhikooli noorema vanusegrupi nõrgemate õpilaste jutukesed võivad koosnedagi vaid paarikümnest sõnast. Seega võib algklassiõpilaste tekstide kogumisel olla tähtis, et korpusesse sisestatavatele tekstidele ei määrataks sõnade alampiiri või et alampiir arvestaks ka vähevõimekate õppijatega.

Tekstide autentsus on õppijakorpuse puhul problemaatiline, kuna õppijad on enamasti situatsioonis, kus nad on sunnitud teksti looma (Nesselhauf 2005: 40). See olukord ei ole samastatav spontaanse kirjutamise või kõnelemisega, vaid sõltub hinnatavusest, etteantud teemast, ajalimiidist, täidetavast ülesandest jm. Seega võib öelda, et õppijakeele korpuses on tekstid erineva autentsuse astmega – alates inimestevahelisest igapäevasest vabast suhtlusest kuni autentsuseni, mida võimaldavad õppetegevused (Granger 2007: 49). Korpuses peaksid erineva autentsusastmega tekstid olema kogutud eraldi (või vastavalt markeeritud).

Õppijakorpuste koostaja ning uurija Silviane Grangeri väitel (2007: 49–50) ei saa kasutada terminit *korpus* vigastele õppijate tekstidest väljanopitud kontekstita sõnade/lausete kogule viidates, kuna vigadele vastukaaluks on vaja ka korrektseid sõnu/lauseid. Et sisestatud tekst saaks kvalifitseeruda õppijakorpuse tekstiks, peab see olema terviklik diskursus, mis sisaldab nii korrektseid kui vigaseid lauseid (Granger 2007: 62). Terviktekstide põhjal on võimalik teha erinevaid statistilisi analüüse või uurida eraldi tekstiosi – näiteks sissejuhatusi, kokkuvõtteid vm. Vajadusel on võimalik mittevajalik tekstiosa tähelepanuta jätta või korpuse tekstide seast oma vajadustele vastav valik teha (Bowker, Pearson 2002: 10).

Teatud juhul, kui korpuse maht on piiratud ning koostamise eesmärgiks on näiteks interaktiivse õppeprogrammi jaoks vigaste keelendite kogumine, võib korpus siiski koosneda ka ainult vigastest lausetest ning nende parandustest.

Sinclair (1995: 18) väidab, et kuna keel areneb kogu aeg, ei saa korpus kunagi valmis, vaid peab aina kasvama. Õppijakorpus võiks olla avatud struktuuriga, et sellele oleks võimalik vajadusel tekste lisada.

Korpus võib olla tavalise lihtsa teksti kujul, kuid sageli on vaja tekstile lisada mingit infot ehk korpust märgendada. Märgendamisest tuleb pikemalt juttu 4. peatükis.

1.2. Korpuse representatiivsus

Kogu keelekasutust ei ole võimalik fikseerida ja seega tuleb korpuse tarvis valida keelest mingi läbilõige, mis esindaks erinevate valdkondade keelekasutust õiges proportsioonis. Suured korpused võivad sisaldada üsna juhuslikult kogutud materjale, näiteks erinevatelt alustelt valitud tekste, mis sobivad keele üldiste seaduspärasuste kvalitatiivseks uurimiseks. Allkeele uurimiseks mõeldud tekstivalik peab olema hästi läbi mõeldud ning eesmärgipäraselt kogutud, kuna erikeele korpus peab andma teatud keelevaldkonna kohta üldpildi ehk olema selle valdkonna suhtes representatiivne. Representatiivsus on eriti oluline suletud korpuste puhul, kui korpuses sisalduvat ei saa hiljem enam muuta. (McEnery, Wilson 1997: 21–23)

Ideaalis sisaldab representatiivne korpus kõiki vastava perioodi tekstiliike, millega inimesed uuritaval perioodil kokku puutusid. Euroopa keeleõppe raamdokument (EKR 2007: 113) esitab kirjaliku teksti liikidena näiteks raamatud (ilu- ja teadusajakirjandus, kirjandusajakirjad); ajakirjad, ajalehed, käsiraamatud, õpikud, koomiksid, brosüürid ja tutvustavad trükised, reklaamlehed, reklaamtekstid, avalikud sildid ja teated, kaupade pakendid ja etiketid, piletid, blanketid jm küsimustikud, sõnaraamatud, tesaurused, äri- ja ametikirjad, faksid, isiklikud kirjad, kirjandid, harjutused, memod, aruanded jm dokumendid, märkmed ja sõnumid, andmebaasid (uudised, kirjandusteosed, üldteave jms) (EKR 2007: 113). Representatiivses korpuses peaksid kõik kasutusalad võrdselt esindatud olema. Täielik representatiivsus ei ole reaalselt saavutatav (Muischnek 2006a).

Douglas Biberi (1996: 177) koostatud representatiivsel õpijakeele uurimiseks mõeldud korpusel *Corpus of elementary student speech and writing* (vt tabel 1) on kolm osa – lastele kirjutatud tekstid (31 000 sõna), laste poolt kirjutatud tekstid (14 000 sõna) ja laste kõnekeel (17 000 sõna).

Tabel 1. D. Biberi loodud representatiivne õppijakeele korpus (Biber *et al.* 2006: 176)

laste jaoks kirjutatud tekstid	**tekstide hulk**	**sõnade arv**
lihtsustatud tekstid (ingl *basal readers*)	23	5 000
reaalainete õppetekstid (ingl *science textbooks*)	25	5 000
ühiskonnateadused (ingl *social studies*)	24	5 000
ilukirjandus	53	16 000
laste kirjutatud tekstid		
õpilastööd	179	14 000
laste suuline kõne		
dialoogid	21	9 000
monoloog	14	8 000
kokku	339	62 000

Saadud korpust ei pea D. Biber eriti heaks, kuna puudub laste keelekasutuses tavaline mängulises situatsioonis kasutatav keel ning teisedki valdkonnad on esindatud vaid väga väikeste mahtudega, mistõttu uuringutulemused ei pruugi olla täiesti usaldatavad (Biber *et al.* 1996: 177). Ometi annab loodud korpus ettekujutuse, missugune võiks olla õppijakeele uurimiseks mõeldud representatiivse korpuse struktuur.

Järgmine peatükk kirjeldab, missugustelt alustelt võib tekstide kogumisel lähtuda. Kirjeldused analüüsivad tegureid, mida tuleb arvestada õppijate tekste sisaldava korpuse loomisel.

2. Tekstivaliku põhimõtted

Korpuse põhjal on võimalik analüüsida ainult seda, mida sisaldavad temasse sisestatud tekstid. Tekstinäited võivad teatud uurimuse läbiviimiseks sobida paremini ja teise jaoks üldse mitte, kuid ükski korpus ei ole iseenesest hea või halb (Lemnitzer, Zinsmeister 2007: 103). Korpusesse valitavate tekstide karakteristikuid on palju ning sõltuvalt korpuse eesmärgist on neil erinev kaal. Õppijakorpuse koostamisel võib lähtuda õppijast, ülesandest, kogumise viisist või teksti enda karakteristikutest. Valitud lähtealus võib mõjutada tekstis esinevate vigade hulka.

2.1. Teksti karakteristikud

Teksti põhjal võib õppijakorpusi jagada meedia, keele, teema või žanri järgi.

- **Meedia**

Korpus võib esineda kas kirjaliku, suulise või segakorpusena. Kirjaliku keele korpused toetavad eelkõige kirjutamis- ja lugemisoskuse uurimist ning arendamist (Eslon, Metslang 2007: 101). Kirjalikke tekste on kergem koguda ning valdav osa korpusi peegeldabki kirjalikku keelekasutust. Suulise kõne analüüsimiseks tuleb vestlus lindistada ja lindistused transkribeerida, mis teeb suulise kõne korpuse loomise aja- ja töömahukaks. Ühes tunniajalises lindistuses on transkribeeritult u 15 000 sõna. Tänapäeval lisatakse suulise kõne korpusesse ka videoklippe. (Keeffe *et al*. 2007: 2)

Eesti kõnekeele korpust haldab TÜ suulise kõne uurimisrühm. Õppijakeele ja kooliealiste laste keelekasutust ei ole spetsiaalselt kogutud, kuid uurimisrühma käsutuses on siiski mõned koolitundide lindistused ning logopeediliste erivajadustega laste ja nende õpetajate vestlused. Ka õpilased on emakeeleolümpiaadile saatnud transkribeeringuid, kuid need ei vasta korpuse koostamisel esitatud nõuetele (näiteks on transkribeering, kuid puudub helikassett vms) (Gerassimenko 2008, eravestlus).

- **Keel**

Ükskeelsed korpused sisaldavad enamasti emakeelekõnelejate suulist või kirjalikku keelekasutust. Kaks- või mitmekeelsed korpused ehk paralleelkorpused või võrdluskorpused võivad sisaldada näiteks ühest keelest teise tõlgitud tekste (ingliskeelne novell hispaania keelde tõlgituna ja hispaaniakeelne novell inglise keelde

tõlgituna) või tekste, mis on kirjutatud samal ajal, samas žanris, aga mitmes erinevas keeles (ka murdekeeles) või registris. Eesti keeles on olemas inglise-eesti ja eesti-inglise euroseadusandluse paralleelkorpus ning George Orwelli "1984", mille paralleeltekstid on bulgaaria, inglise, rumeenia, sloveeni, tšehhi ja ungari keeles.

Paralleeltekst võib olla originaaliga seotud kaude – tekstiosadele, lõikudele ja lausetele viitavalt või otse – iga tõlketeksti lause juurde on kirjutatud originaali lause number (Kaalep 1999: 17). Kakskeelne paralleelkorpus sobib tõlkijate õpetamiseks, tõlkimisel terminoloogia ja stiili valimiseks (Hunston 2002: 14–16).

Õppijakorpuste kontekstis tähendab paralleelkorpus erineva emakeelega õppijate võrreldavatelt alustelt koostatud korpust või (ema)keeleõppija kirjutatud tekste ja samade tekstide paranduste ning kommentaaridega versiooni. Kui ühte teksti parandab mitu inimest, on võimalik võrrelda ning analüüsida ka parandusi. Eesti keelt võõrkeelena õppivate tudengite keelekasutust koostatakse TÜ eesti keele võõrkeelena osakonnas.

- **Teema ja žanr**

Sõltuvalt korpuse eesmärgist võib korpus sisaldada kõiki teksti tüüpe kas eraldi või teatud valimina. Üldkeele korpuseks (e koondkorpuseks[1]). nimetatakse erinevat tüüpi tekstide korpust, milles sisalduvate tekstide variatiivsus on suur. Eesti keele koondkorpus sisaldab suures mahus ajalehetekste ja väiksemas mahus muid tekste – Riigikogu stenogramme, Eesti ja Euroopa seadusi, ilukirjandust ja jutustavaid tekste.

Erikeele korpuse tekstid on spetsiifilised: seotud kindla teema või tekstitüübiga. Tekstide valikukriteeriumitest sõltub korpuse sisu. Õppijate loodud kirjalikud tekstid võib jagada etteantud teemast lähtuvateks esseelaadseteks, analüüsivateks või arutlevateks õppetekstideks; uurimuslikeks tekstideks ja loomingulisteks (ilu)kirjanduslikeks tekstideks. Enamasti esindavad õpilastekstide kogud vaid ühte teksti tüüpi – arutlevaid esseelaadseid kirjandeid – ning on oht, et analüüsitulemused ei iseloomusta mitte õppijatekstide eripära, vaid pigem konkreetse žanri keelekasutust (Barlow 2005: 336).

[1] Ka *kontrollkorpus* (Lemnizer 2006: 106) või *referentskorpus* (Hunston 2002: 14–17)

Õppijakorpuste loomisel on peetud oluliseks, kas õppija kirjutis lähtub kirjandusteosest või on vabateemaline. Osa uurijaid on kirjandusteoseid analüüsivad esseed õpilastekstide korpusest välja jätnud (Nesselhauf 2005: 45).

Teksti kirjutamist mõjutavad ka teema ning žanri tuntus. Õppija jaoks on veaohtlikumad võõral teemal või tundmatus žanris kirjutatavad tekstid, samuti sisu poolest ebahuvitavad tekstid. Ka kirjalikke tekste on peetud suulistest veaohtlikumaks (vt tabel 2). (Ellis, Barkhuizen 2005: 56)

Erikeele korpustega seoses võib rääkida pedagoogilisest korpusest, millesse kuulub õppijate jaoks kirjutatud keel ehk sisendkeel. Pedagoogiline korpus koosneb paljudest väikestest **erikeele korpustest** – näiteks õpikutekstidest, populaarteaduslikest ajakirjandustekstidest, ilukirjandustekstidest, intervjuudest, õppijate referaatidest, aga ka filmisubtiitritest, reklaamtekstidest, hobiartiklitest, lasteajakirjades ilmunud kirjutistest jm (Hunston 2002: 16). Erikeele korpuse tekstid sobivad õppija sisendkeele analüüsimiseks, kuid on samal ajal ka õppijatele eeskujuks konkreetses keeleregistris kasutatavate väljendite ja stiilierinevuste õppimisel. Pedagoogiliste korpuse tekste võib võrrelda üldkeelekorpusega, et näha, kas õppijate sisendkeel on samasugune nagu üldkorpuses talletatud keel, või õppijate kirjutistega, et näha, kuivõrd sisendkeel õppijate keelt mõjutab või mil moel sisendkeeles eristuv õppijakeeles kajastub. Näiteks tuleneb parema emakeeleoskusega õpilaste kirjandeid üldiseloomustav kantseliitlik keelekasutus tõenäoliselt ajakirjanduse ning muu meedia kaudu saadud keelesisendist.

Douglas Biberi (2007: 175–177) koostatud emakeeleõppijate mikrokorpuse *Corpus of elementary student speech and writing* koostises on sisendkeele tekstidel oluline osa. Lastele kirjutatud tekstid moodustasid poole kogu korpusest (31 000 sõna).

Tabel 2. Teksti veaohtlikkust mõjutavad tegurid keelest lähtuvalt (koostatud Ellis, Barkhuizen 2005: 56 põhjal)

keelest lähtuvad mõjurid	**veaohtlikum**	**mitte nii veaohtlik**
keel	võõrkeelena õpitav keel	emakeel
meedia	kirjalik	suuline
teema	õppija jaoks võõras õppija jaoks ebahuvitav	õppijale tuttav õppijale huvipakkuv
žanr	õppija jaoks võõras	õppijale tuttav
teksti tuntus	õppija jaoks võõras	õppijale tuttav

2.2. Autorist lähtuv tekstivalik

Enamasti koostatakse korpused erinevate autorite tekstidest, kuid on võimalik luua ka ühe autori tekstil põhinevat korpust. Tekstides sisalduvat keelekasutust mõjutavad teksti autori (õppija) vanus, sugu, haridustase, keelekeskkond, õppija emakeel ja päritolu, sotsiaalne kuuluvus, teised õpitavad keeled ning nende osaoskuste tase, varemõpitud keeled ja keele prestiižsus. (Barlow 2005: 337; Ellis, Barkhuizen 2005: 56) Arvestada võib ka kognitiivseid (õpistiil) või afektiivseid (motivatsioon, suhtumine) tegureid (Tono 2002: 1). Suulise kõne tekstide kogumisel on soovitatud arvestada kõnelejate hulka, sugu, vanust, sünnikohta, ametit, haridustaset jm (Keeffe *et al*. 2007: 7). Uurija peab nimetatud tegurite seast valima need, mis on tema uuringu seisukohalt olulised ning mida saadav keelematerjal lisada võimaldab.

Kirjeldatu eeskujul võib loetleda teksti veaohtlikkust mõjutavad tegurid õppijast lähtuvat (vt tabel 3).

Tabel 3. Teksti veaohtlikkust mõjutavad tegurid õppijast lähtuvalt (koostatud Ellis, Barkhuizen 2005: 54–57 põhjal).

teksti autorist lähtuvad mõjurid	
vanus	vanemad inimesed on keelemuutustele vähem vastuvõtlikud, noored vastuvõtlikumad
haridustase	Emakeeleõppijate keeleoskustasemete kirjeldusi Eestis loodud ei ole. Õppijate emakeeleoskuse tase on määratav klassi, ainekava ning keskmise emakeelehinde abil. Kõrgharidusega emakeelekasutajate keelt peetakse kõrgtasemeliseks.
keelekeskkond	naabrid, sõbrad, kodune keel, vaba aja tegevuste keel (arvutimängud, kaardimängud, filmid vm).
teised õpitavad keeled	esimene ja teine võõrkeel koolis
emakeele prestiižsus ja üldine suhtumine keelde	meedia keelekasutus, registritest kinnipidamine avalikus keelekasutuses, suhtumine korrektsesse keelekasutusse

2.3. Situatsioonist lähtuv tekstivalik

Kui korpusesse kogutakse tekste lähtuvalt nende loomise situatsioonist, tuleb arvestada õppijale antud ülesannet, selle kirjutamise eesmärki ja kontrollitavust, kirjutamise spontaansust, abimaterjalide kasutamise võimalust ning ajapiirangut: teksti kirjutamise aeg võib olla piiritlemata (näiteks koduse töö puhul) või piiratud (Tono

2002: 1; Barlow 2005: 337). Suulise kõne uuringute puhul on soovitatud arvestada ka vestluspartnerit (Barlow 2005: 337).

Õppijatele kirjutatud tekstide puhul on oluline teksti kirjutamise aeg, trükistel esmane väljaandmisaeg (Sinclair 1995: 18).

Teksti esitamise situatsioonist võib sõltuda ka vigade hulk. Spontaanne esitus on veaohtlikum kui ettevalmistatud esitamine. (Ellis, Barkhuizen 2005: 56)

2.4. Kogumisviisist lähtuv tekstivalik

Sünkroonilised korpused sisaldavad ühel ajal kirjutatud tekste. Diakrooniliste ehk ajalooliste korpuste tekstid on kogutud teatud ajajärgust lähtudes ning võimaldavad uurida keeles aja jooksul toimuvaid muutusi. Eesti keele ajaloolisteks korpusteks on vana kirjakeele korpus ja Eesti kirjakeele korpus 1890–1990.

Õpilastekstide diakrooniline korpus võimaldab näiteks viimase viie aasta jooksul kirjutatud ning riiklike eksamikirjandite kirjutamise algusaegadel loodud tekste võrreldes analüüsida, missugused keelekasutuse jooned on tingitud konkreetse perioodi sisendkeelest ning missugused on üldisemad, õppija keelele üldomased.

On korpusi, mis esitlevad õppija keelt arengus (ingl *longitudinal*) so pikema aja jooksul – kuude või aastate kaupa. Selline on näiteks Houseni suulise kõne korpus *Corpus of Young Learner Interlanguage*, mis sisaldab Euroopa koolides õppivate erineva keeleoskustaseme ja erineva emakeelega laste ingliskeelseid tekste. Enamasti ei ole materjali kogumiseks mitmeid aastaid aega. Kui uurija soovib keeleoskuste arenemise kohta andmeid, on tavaks koguda need homogeensetelt erineva tasemega õppijate gruppidelt (ingl *quasilongitudinal*). (Granger 2007: 51)

Parajasti toimuvate keelemuutuste uurimiseks sobib seirekorpus ehk monitorkorpus. Seirekorpuse tekste sisestatakse kindlate vaheaegade järel. Analüüsitavate tekstide hulk on konstantne, ning omavahel hästi võrreldav. Seirekorpuse materjaliks sobivad ajakirjandustekstid. (McEnery, Wilson 1997: 22; Hunston 2002: 14–16; 31) Õppijakeele muutuste jälgimiseks võib kasutada õpilasajalehti, almanahhe, eksamikirjandeid, kirjandivõistlustele saadetud töid või tasemetöid.

Tabel 4 koondab need tegurid, millest sõltub kogutavate tekstide spetsiifilisus.

Tabel 4. Tegurid, mis muudavad korpuse tekstikogu spetsiifiliseks (Ellis, Barkhuizen (2005); Granger (2007), McEnery, Wilson (1997), Keeffe *et al.* (2007) põhjal)

tekst	meedia, teksti teema, teksti tüüp, tekstitüübi tuntus, keel
autor	vanus, sugu, haridustase, emakeel, keelekeskkond, sotsiaalne staatus, teised õpitavad keeled, õpitava keele prestiižsus.
situatsioon	ülesanne ja selle eesmärk, kontrollitavus, spontaansus, abimaterjalide kasutamise võimalus, ajapiirang
kogumisviis	sünkrooniline, diakrooniline, *longitudinal*, *guasilongitudinal*

2.5. Metaandmed

Tabelis 4 loetletud teguritest sõltub õppija keelekasutus ning korpuse tekstidele lisatavad metaandmed peavad keelematerjali kogumisel arvesse võetud tegureid kajastama. Lisatavate metaandmete ulatus tuleb enne tekstide kogumist ära otsustada, sest hiljem on andmeid väga raske, kui mitte võimatu täiendada. Korpuses sisalduvatele tekstidele lisatud metainfo ulatusest sõltub sotsiolingvistiliste uuringute võimalikkus.

J. Sinclairi (1995: 20) arvates peab autori kohta teada olema tema vanus, sugu ja emakeel ning korpuse teksti juurde kuuluv info peab kandma teavet selle kohta, kas lisatud tekst on ilukirjanduslik või mitteilukirjanduslik; raamatutekst või ajakirjandustekst; ametlik või mitteametlik (Sinclair 1995: 20).

1990. aastatel Louvaini ülikoolis Grangeri (Barlow 2005: 339; Granger: 2006) koostatud tuntuimas õppijakorpuses ICLE – *International Corpus of Learner English* on tekstile metaandmetena lisatud ülesande kohta selle tüüp (*task*) ja meedia (suuline v kirjalik), teksti kohta selle teema, žanr, pikkus ja pealkiri ning õppija kohta tema identifitseerimiskood, keeleoskuse tase (*profiency level*), emakeel, vanus, amet (sotsiaalne staatus) ning õppimise kontekst (ingl *learning context*) (vt joonis 1). Hiljem on õppijate andmeid lihtne võrrelda (vt joonis 2).

Tag format

```
<student id = "028">
<sex>male</sex>
<proficiency>intermediate</proficiency>
<language>Dutch</language>
<age>16</age>
<occupation>student</occupation>
<learning_context>school</learning_context>
</student>
```

Joonis 1. Õppijakorpuses ICLE õppija kohta lisatud andmed (Ellis, Barkhuizen 2005 : 338)

Database format

ID	Sex	Proficiency	Mother tongue	Occupation	Age	Learning Context
028	male	intermediate	Dutch	Student	16	school
029	female	intermediate	French	Student	17	school
030	male	advanced	French	Student	18	school
031	female	intermediate	Dutch	Student	17	school

Joonis 2. Õppijakorpuses ICLE sisalduvate tekstiautorite metaandmed (Ellis, Barkhuizen 2005: 338)

Tallinna ülikooli vahekeele korpuse informantide kohta on kogutud järgmised andmed: sugu, vanus, elukoht, sotsiaalne staatus (teenistuja vm), emakeel, kodukeel, haridus, keeleoskuse tase ning teksti kohta selle maht, liik (essee, kiri, harjutus vm), teksti koostamise laad (koostatud abivahendeid kasutades või ilma) (Eslon 2007: 100-101). Tartu ülikooli eesti keele (võõrkeelena) õppetoolis koostatav paralleelkorpus sisaldab informantide kohta teavet nende emakeele ja keeleoskustaseme kohta.

Korpuse kättesaadavakstegemine väljaspool kitsast tegijateringi nõuab lisapingutusi formaalse ühtluse ning dokumentatsiooni loomisel, mida tuleks korpuse tegemisel algusest peale arvestada (Kaalep 1999: 20). Eestis kehtivate seaduste kohaselt peab terviktekstide esitlemiseks olema nende autorilt kirjalik nõusolek (alaealise puhul tema vanemalt või hooldajalt), tekstiosi on õigus kasutada ka nende autorile viitamata.

3. Korpus keeleõppes

Keeleuurijate ja -õppijate huvid korpuse kasutamisel on mõnevõrra erinevad. Mõlemad sihtgrupid tahavad küll saada ettekujutust, kuidas keel funktsioneerib, kuidas teatud sõna, väljendit või grammatilist konstruktsiooni kasutatakse, kuid uurija eesmärgiks on vormi kirjeldada ja teha üldistusi, keeleõppija soovib teha järeldusi ainult iseenda jaoks ning keelt õigesti kasutada, isegi kui ta ei saa täpselt aru reeglitest, mis keelt juhivad. (Willis, Willis 2007: 188–189)

Keeleõppijatele, õpetajatele ning õppematerjalidega tegelejatele on korpusuuringud hakanud huvi pakkuma alles viimasel paarikümnel aastal. (Ranalli 2003: 2–3) Üks osa teadlasi on seisukohal, et korpuste analüüs on õppeaine sisu, meetodite ja õppematerjalide arendamiseks oluline (Sinclair 1986; Johns 2007, Mc Enery 1996; Bernardiny 2004; Tribble 1997, Kettemann 2000; Bick 2004), teised kõhklevad, kuivõrd ja kas korpuste kirjeldused üldse õppijaid kuidagi aitavad. Teadlased, kelle arvates on korpusuuringud õppetöös olulised, esindavad omakorda kahte erinevat arvamust selle kohta, mil määral võib korpus keeleõpet mõjutada. Esimese, nö mõõdukama lähenemise esindajad väidavad, et korpus aitab õpetust kujundada (*modeling*) ja teine, radikaalne korpusepõhine lähenemine (*corpus-driven approach*) peab korpust ainsaks õigeks ja autentseks õppematerjaliks. (Ranalli 2003: 3–6)

Korpuste kasutamist keeleõppes toetavad peamiselt kahte liiki väited, millest üks osa rõhutab intuitsioonil põhineva õpetuse vähenemist ja reaalse keelekasutuse suurenemist (Mc Enery, Wilson 1997: 116), teised hindavad õpetuse muutumist uurimuslikuks, induktiivsks (Bernardini 2007: 104) ning õppija muutumist teadlikuks ja aktiivseks (Bowker, Pearson 2002: 1, Ranalli 2003: 10). Meunier (2007: 27) rõhutab, et õppijakorpuse abil on võimalik leida õppijate jaoks raskusi tekitavaid keelendeid ja saadud teavet kasutada õppekava koostamise lähtealusena. Õpilase enda formuleeritud keeleteadmised on märksa püsivamad pähetambitud tõdedest (Lamy *et al.* 2007: 11, Bernardini 2007: 118, Ranalli 2003: 13, Johns 2007: 5).

Vastuväidetena on kritiseeritud korpuse näitelauseid kui õppijate jaoks ebahuvitavaid, redigeerimata (Lamy *et al.* 2007: 12), liiga ulatuslikku tõlgendamist vajavaid (Tribble 1997: 2) ning äärmiselt keerulisi ja seetõttu keeleõppesse

mittesobivaid (Ranalli 2003: 9). Arvatakse, et nii õpilased (Meunier 2007: 37) kui õpetajad (Lamy *et al.* 2007: 12) saavad end reeglite kaudu keelt õppides palju kindlamalt tunda. Vaidlusi on tekitanud ka tekstide tegelik autentsus, sest näited esitatakse ju tekstuaalse ja sotsiaalse kontekstita (Ranalli 2003: 8).

Lemnitzer (2007: 167) väidab, et korpuspõhise uurimisviisiga saavad hakkama vaid täiskasvanud õppijad või üliõpilased, mitmed teised asjatundjad näiteks Hunston (2002: 171) Bernardini (2007: 118) ja Lamy (2007: 5) on aga seisukohal, et ka nooremad õpilased suudavad hea juhendamise korral keelele induktiivselt läheneda.

3.1. Korpuspõhised õppematerjalid

Uurimine ja keele õpetamine on ühendatud Lõuna-Taani ülikoolis, kus keelekorpuse teaduslike eesmärkide – märgenduse automatiseerimise ja korpustest päringute tegemise kõrvalharuks on korpusepõhine keeleõppeprogramm *Visual Interactive Syntax Learning* (VISL) Morfoloogiliselt ja süntaktiliselt märgendatud korpust kasutatakse veebipõhiste interaktiivsete grammatikaülesannete loomiseks. VISL koondab lisaks taani, inglise ja saksa keelele veel 22 keele struktuuri õpetamiseks mõeldud ülesandeid. Mängulised harjutused on loodud käsitsi märgendatud autentsete korpuste põhjal. Süsteem põhineb kategoriaalgrammatilisel parseril, mis esitab laused sõltuvusanalüüsi abil. Teiste keelte seas on 100-lauselise märgendatud korpusega esindatud ka eesti keel (projektijuht Heli Uibo). (Bick, Uibo 2005: 195–201; Lemnitzer 2007: 167–168) VISLi eestikeelne rakendus on alles väikesemahuline. Õpiprogrammi kasutamist ei soosi ka inglise keele struktuurile loodud keskkond, kuid tegu on huvitava projektiga. Taolised interaktiivsed õppevahendid väärivad arendamiseks nii suuremat rahalist kui ka inimressurssi.

Spetsiaalseid erikeelt sisaldavad sisendkeele korpusi võib rakendada õppe-eesmärgil **eeskujukorpus**tena. Eeskujukorpuste mõte on pakkuda erinevaid teemasid käsitlevate tekstide eeskujusid – ärikirjadest, arvustustest, eksperimendikirjeldustest ja avaldustest kuni reisikirjade ja hobitekstideni. Erialatekstid erinevad sõnavara, ülesehituse ning kasutatavate grammatiliste ja süntaktiliste struktuuride kasutussageduse poolest. Tekst, mis on korrektne üldkeele grammatikareeglite kohaselt, ei pruugi vastata erikeele nõuetele terminoloogia või stiili poolest (Bowker, Pearson 2002: 27; 36). Tribble

(1997: 3–5) soovitab eeskujukorpuse tekstidena kasutada näiteks teatmeteoseid (žanriteks kirjeldused, bibliograafilised tekstid ja selgitused).

Mõjukaim korpusenäidete põhjal loodud võõrkeeleõpikute sari on **Collins Cobuild**, mis sai alguse 1980. aastal Birminghami ülikoolis John Sinclairi juhitud projektist (*Collins Birmingham University International Language Database* COBUILD) (Keeffe *et al*. 2007: 17; Sinclair 1991: 2). Cobuildi materjalid põhinevad kõik inglise keele korpusel (*Bank of Englisch*), mis sisaldab mitmeid miljoneid inglise üldkeele tekste ja suhtluskeelt. 2007. aastaks on välja antud 16 korpuse materjalide põhjal koostatud õpikut (sõnaraamatud, grammatikaõpikud (näiteks Willis 1991)). Eriti oluliseks peetakse, (McEnery *et al.* 1997: 11) et õpikutes on viidatud sõnade struktuuri, kasutamise ja tähenduse vahelistele seostele. Korpuses sisalduv keelematerjal on põhjalikult analüüsitud ning sellele tuginevad grammatika õpetamise materjalid jagatud kolme kategooriasse:

1) viitav (teatme) grammatika (ingl *reference grammars*) korpuse *Bank of Englisch* empiirilisel analüüsil põhinevad grammatikakirjeldused, mis on suunatud spetsiaalselt keeleõppijale (ka iseõppijale) ning sisaldavad peale kirjelduste ka korpuspõhiseid harjutusi.

2) *guides to individual areas of grammar* – ühele grammatika teemale keskendunud õpikud, mis esitlevad keelemustreid. Õpikutesse on lisatud autentseid näiteid ja harjutusi.

3) andmete raamatud (ingl *data books*) – esitlevad märksõnade (KIWIC) konkordantse ja suunavad õppijat valitud teema kohta kogutud töötlemata keelematerjali uurima ning analüüsima (vt pt 4.1.2.). (McEnery *et al*. 1997: 11)

COBUILDi projekt on oluliselt mõjutanud inglise keele grammatika õpetamist inglise keelt kõnelevates maades. Õpetajad on väitnud, et nüüd, pärast aastatepikkust keskendumist suhtlusõpetusele, julgevad nad taas rääkida ka grammatika õpetamise vajalikkusest (McEnery *et al.* 1997: 8). Cobuildi õpikud lähenevad grammatikale keelemustreid esitledes ja nimetavad grammatikat sõnavara kasutajaliideseks (Keeffe *et al.* 2007: 17).

COBUILDi õpikud esindavad õppimise uurimuslikku induktiivset suunda ning lähtuvad 21. sajandil populaarsust kogunud Jean Piaget´ konstruktivistlikust õpiteooriast, mille ideed sobivad hästi korpuspõhise analüüsi põhimõtetega. Konstruktivistlik õpikäsitluses lähtub seisukohast, et "ei ole olemas sõltumatut ja objektiivset reaalsust. Tõde ja teadmised ei eksisteeri väljaspool inimmõistust, kuna kõik, mida me teame, on meie endi poolt meie oma mõtetes, oma keelelises ja ajaloolises perspektiivis ühiskondlikult konstrueeritud. Lisaks tähenduse loomisele peab õppija looma ka tähenduste süsteemi – iga uus loodav mõte peab sobituma meie olemasolevate teadmistega. Uue õppimiseks peame omama mingit olemasolevat eelteadmist ja struktuuri, millele uus teadmine rajada." (Probleemõppe lähtekohad 2007)

David. H. Jonassen (1994: 34–37) toob esile kolm konstruktivistide jaoks olulist lähtepunkti

1) autentne kontekst (*context*) - õppimiseks kasutatakse autentseid (tegelikul keelekasutusel põhinevaid) õppematerjale. Algmaterjali ei tohi lihtsustada ega näiteid spetsiaalselt õppeülesannete tarvis luua;

2) teadmiste konstrueerimine (*construction*) – ennustavate, seletavate, järeldavate mõttemudelite (mental models) loomine ning nende analüüsimine; õppija võimalus aktiivselt teadmist konstrueerida (luua), ideedega eksperimenteerida ja loodu kohta tagasisidet saada.

3) meeskonnatöö (*collaboration*). Tulemuseni jõudmine nõuab meeskonnatööd ning oskust lähtuda erinevatest vaatenurkadest.

Silvia Bernardini (2007: 102–123) lisab veel kaks olulist punkti

4) Konstruktivistlik õpikäsitlus näeb õpetajat pigem mentori kui kõigeteadja rollis.

5) Õppijad muutuvad iseseisvamaks ja vastutavamaks, iseseisvus ja vastutus omakorda tõstavad motivatsiooni.

Kognitiivse teooriate järgi (*Concept to classroom*. 3.01.2008) on õppimise põhimotiiviks õppija enda sisemine huvi ja sellest tulenev aktiivsus. Teadmine on

indiviidi meeltes mentaalse konstruktsiooni tulemusel tekkinu: kontekst ja situatsioon määravad teadmuse. Uusi sõnu ei saa õppida sõnastiku järgi, sest sõnastikus on terminid nende kasutamise kontekstist välja rebitud. Mitte üksnes asesõnad, vaid enamus sõnu on nn deiktilise ehk indekseeritud tähendusega (*indexed meaning*), s.t. kontekstist välja võttes kaotavad nad esialgse tähenduse. Uute terminite sisuline mõistmine ja tähendusseosed tekivad senitundmatu kasutamisel konkreetses kontekstis ja situatsioonis. (Probleemõppe lähtekohad 2007)

"Ehkki keeleõpe arvuti abil ei puuduta otseselt /.../ keeletehnoloogiat, aitab ta siiski laiendada /.../ keele elektroonilist kasutussfääri ja seega toetab ka keeletehnoloogiat" (Muischnek *et al.* 2003: 71). Korpuse põhjal koostatud materjalid toetavad uurimuslikku õppimist ja sunnivad õppijat kontsentreerima oma tähelepanu teatud probleemi lahendamisele ja seaduspära tabamisele. Töö lisas esitatakse mõned eesti keele üldkorpuse põhjal loodud näidisharjutused eesti keele grammatika (lisad 1, 2) ja sõnavara (lisa 3) õppimiseks.

Korpuspõhine õpe toob koolitundi teistsuguse lähenemise, täiendab teisi õpetamismeetodeid (Meunier 2007: 37), kuid kindlasti ei sobi kõikide teemade õpetamiseks.

3.2. Õppijakeelt sisaldavad korpused

Kuigi korpusepõhiseid lingvistilisi uuringuid keeleõppe eesmärgil peetakse oluliseks (Kitsnik 2006: 94–98), on emakeeleõppijate, eriti laste ja noorte keele uurimiseks kogu maailmas vähe spetsiaalseid korpuseid (Biber *et al.* 2006: 175). Olemasolevatele korpustele ligipääs on sageli piiratud. Seniste õppijakorpuste üheks ühiseks jooneks on, et need sisaldavad kindlasti teise keele või võõrkeeleõppija tekste/kõnesalvestusi, millele on võrdluseks lisatud ka samaealiste emakeelekõnelejate tekste.

Küsimusele, miks korpusepõhistel õppijakeele uuringutel eelistatakse keskenduda mitte-emakeelsele õppijale ega tähtsustata emakeeleõppija keelekasutuse uurimist, ei ole töö kirjutajal õnnestunud veenvat selgitust leida. Ehk on põhjus selles, et arvutipõhised õppijakorpused (ingl *computer learner corpora*) on rakenduslingvistikas alles noor (Granger 2006) ning sageli kommertsliku eesmärgiga ala.

Järgnevalt annan ülevaate olemasolevatest laste- või õppijakeelt sisaldavatest korpustest.

Belgia

Louvain Corpus of Native English Essays (LOCNESS[2]) sisaldab nii inglise keelt emakeelena kui ka teise keelena õppijate kirjutatud keele tekste. Korpuses on

Biti õpilaste A taseme esseed	60 209 sõna
Briti üliõpilaste esseed	95 695 sõna
Ameerika ülikoolide tudengiesseed	168 400 sõna
kokku	324 304 sõna

Kirjutajad on alates vanusest 17 aastat. Korpus sisaldab ka inglise keelt teise keelena õppijate tekste.

International Corpus of Learner Englisch (ICLE[3]) on loodud 1990. aastatel Belgias Louvaini ülikoolis ning sisaldab u 3 miljonit sõna ingliskeelseid tekste 14 erineva emakeelega (hiina, tšehhi, saksa, hollandi, soome, prantsuse, saksa, jaapani, poola, vene, hispaania, rootsi) üliõpilastelt, kes õpivad inglise keelt. Allkorpused ei ole suured – u 200 000 sõna. Kontrollkorpuseks on inglise keelt emakeelena õppijate allkorpus. ICLE on paljude teaduslike õppijakeele uuringute andmebaasiks. (Biber *et al.* 2006: 283; Granger: 2006; Barlow 2005: 339)

Louvain International Database of Spoken English Interlanguage (LINDSEI[4]) loodi 1995. a Belgias Louvaini ülikoolis. Korpus sisaldab u 100 000 sõna prantsuse emakeelega inglise keele õppijate tekste, mille võrdlusaluseks on inglise keelt emakeelena rääkijate kirjutised. Projekti on kavas laiendada ning lisada Jaapani, Rootsi, Bulgaaria, Hiina ja Hispaania taustaga õppijate keelenäiteid.

Suurbritannia

Longman Learner Corpus (LLC[5]) sisaldab u 20 miljonit sõna ja *Cambridge Learner Corpus* (CLC[6]) sisaldab u 10 miljonit sõna. Korpused kuuluvad Cambridge ülikooli

[2] http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/locness1.htm

[3] http://cecl.fltr.ucl.ac.be/Cecl-Projects/Icle/icle.htm

[4] http://cecl.fltr.ucl.ac.be/Cecl-Projects/Lindsei/lindsei.htm

[5] http://www.pearsonlongman.com/dictionaries/corpus/learners.html

[6] http://www.cambridge.org/elt/corpus/learner_corpus.htm

juurde ning on kommertsiaalsed. Korpustes märgendatakse grammatikavigu, analüüsitakse õppijate tekste eesmärgiga parandada õppevahendeid, muuta neid õppijate vajadustele vastavalt. Korpused ei ole avalikuks kasutamiseks. (Ellis, Barkhuizen 2005: 241)

The Bergen Corpus of London Teenage Language (COLT[7]) on teismeliste suulise kõne korpus. Korpuse materjale hakati koguma 1993. aastal. Korpus koosneb 62 erineva sotsiaalse taustaga 13–17-aastase Londoni noore spontaansetest vestlustest (lisaks rollimängude ja pildi kirjeldamise lindistused). Salvestised on tehtud aastatel 1994–1995. Vestlused on transkribeeritud, saadud tekst morfoloogiliselt märgendatud. Hiljem on osa korpusest märgendatud ka prosoodiliselt. Võrdlusaluseks on 26 briti lapse lindistused. Sel kujul on COLT liidetud *British National Corpuse* (BNC[8]) hulka. Korpuse uuringutulemuste kohta on kirjutanud J. Benjamin 2002 "*Trends in Teenage Talk*". (Ellis, Barkhuizen 2005: 241)

Suurbritannias Leedsis koostatakse ka korpust nimega *Polytechnic of Wales* (POW[9]), mis sisaldab 61 000 sõna inglise laste suulist keelekasutust. (Biber *et al.* 2006: 284)

Ameerika

Põhja-Arizona ülikoolis on loodud (Grabe ja Reppen) spetsiaalne korpus nimega *Arizona Corpus of Elementary Student Writing*, mis sisaldab u 5 000 kirjandit viiel keeletasemel. Emakeelsete keeleõppijate tekstide kõrval on kogutud ka navajo ja hispaania emakeelega õppijate tekste. Tekstid on kogutud longituudse uuringu käigus. Õppijad kirjutasid mitmel aastal 9 erinevat teksti (arvamus, jutustus, info edastamine jm). Tekstid on kirjutatud abivahendeid kasutamata. (Biber *et al.* 2006: 175–177)

TOEFL 2000 *Spoken and written Academic Language Corpus* (BASE[10]) – T2K-SWAL *Corpus* sisaldab nelja Ameerika Ühendriikide ülikooli nii suulisi kui ka

[7] http://torvald.aksis.uib.no/colt/

[8] http://www.natcorp.ox.ac.uk/what/index.html

[9] http://www.comp.leeds.ac.uk/amalgam/tagsets/pow.html

kirjalikke õppetööga seotud tekste – loengud, õppematerjalid jm. Korpuse suurus on 2,8 miljonit sõna. Korpus ei ole avalikuks kasutamiseks, kuid 2002. aastal on korpuse analüüsi tulemustest kirjutanud Biber, Conrad, Reppen, Byrd ja Helt raamatus "Speaking and writing in the university: A multi-dimensional comparison".

Hiina

Chinese Learner Englisch Corpus (C-LEC[11]) – korpuse projekti juhivad Gui Shichun Guangdong´i ülikoolist ja professor Yang Huizhong Shanghai Jiaotong´i ülikoolist. Korus sisaldab miljon ingliskeelset sõna gümnaasiumiealistelt inglise keele õppijatelt ja tudengitelt. Korpuse tekstides on vead märgendatud. Kokku on 61 veatüüpi (semantilised, leksikaalsed, grammatilised, süntaktilised vead).

Hong-Kong University of Science and Technology (HKUST[12]) inglise keele õppijate keelt sisaldav korpus suurusega u 6 miljonit sõna (avatud). (Ellis, Barkhuizen 2005: 241 andmetel juba 25 miljonit sõna.) Hong Kongi eksamikeskusest kogutud kirjalikud tööd vastavad keeleoskuse A tasemele. (Biber *et al.* 2006: 283)

Rootsi

Uppsala Student Englisch Corpus (USE[13]) – Uppsala ülikooli inglise keele osakonnas õppivate rootsi üliõpilaste esseedest koosnev korpus, mis on koostatud aastatel 1999–2001. Korpus sisaldab 1 489 esseed 440 rootsi inglise keelt õppivatelt tudengitelt (kolmel keeleoskustasemel). Korpuse suurus on 1 221 265 sõna. (http://www.engelska.uu.se/use.html)

Soome keele õppija korpus

Soome keele õppijate eksamitööde kogud asuvad Helsingi ülikooli soome keele ja kirjanduse osakonnas ning sisaldavad soome keelt teise keele või võõrkeelena õppijate tekstide kõrval ka soome keele kui emakeele riigieksamikirjandeid (Eslon 2007: 103). Kogu ei ole avalikuks kasutamiseks.

---

[10] http://www2.warwick.ac.uk/fac/soc/al/research/projects/resources/base

[11] http://langbank.engl.polyu.edu.hk/corpus/clec.html

[12] http://www.ust.hk/eng/about/index.htm

[13] http://www.engelska.uu.se/use.html

Eesti keele õppija korpused

Tallinna ülikooli vahekeelekorpus (EVKK[14]) sisaldab 459 731 sõna eesti keele kui teise keele õppija tekste, mis sisaldab osaliselt tasakaalustatud alg-, kesk- ja kõrgtaseme tekste erineva sotsiaalse taustaga keeleoskajatelt. Korpuse märgendamiseks on kirjutatud spetsiaalne programm. (Eslon 2007: 101)

Tartu ülikooli veakorpus[15] (128 000 sõna) sisaldab eesti keele kui teist keelt õppivate üliõpilaste tekste ja on koostatud paralleelkorpusena. Vigaste lausete kõrvale on loodud korrigeeritud versioon. Vajadusel on lisatud mitu parandusvõimalust.

Norras koostatud õppijakeelekorpused

Kolmas Norrakeelne õppijakorpus *Language learner corpus of Norwegian as a second language* (CONS[16]) sisaldab ainult norra keelt võõrkeelena õppijate keelekasutust.

Saksa keele õppija korpus

Fehler-Annotiertes Linguistisches Korpus (FALKO[17]) Humboldti ülikooli keelekorpus, mis koondab saksa keelt teise keelena õppijate tekste. Uuritakse kakskeelsete õppijate tüüpilisi vigu. Korpuse eesmärk on didaktiline. Korpuse tekstid on osaliselt parandatud, vigade liigitamise ja märgendamise süsteem on loomisel (2006. aasta andmed). Võrdluskorpus sisaldab emakeeleõppijate kirjutisi. (Lemnitzer 2006:117)

Prantsuse keele õppijate korpused

French Interlanguage Database (FRIDA[18]) tekstid koguti aastatel 1998.–2001. prantsuse keelt võõrkeelena õppijatelt, kelle keeleoskus oli vähemalt kesktasemel.

[14] http://evkk.tlu.ee/

[15] http://www.keeletehnoloogia.ee/syndmused/konverents/slaidid/teral.pdf

[16] http://www.ugr.es/~talc6/talc_search/proceedings/60.html

[17] http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/falko

[18] http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Frida/frida.htm

Korpus sisaldab 450 000 sõna, millest u 300 000 sõna on märgendatud. Tekstidele on tehtud vigade statistika ja analüüs. 46 241 viga tuvastati käsitsi ja seejärel lisati veaprogrammi kasutades märgendused. FRIDA alusel on loodud keeleõppeprogramm FreeText, mis suudab parandada õppija vigu ning viidata käsiraamatus vastavale reeglile. Korpus ei ole avalikuks kasutamiseks. (Granger 2003: 470)

French Learner Language Oral Corpora (FLLOC[19]) sisaldab prantsuse keele õppijate suulise kõne lindistusi, mida on transkribeeritud CHILDESi tarkvara kasutades.

Jaapanis loodud keeleõppija korpused

Corpus of Englisch by Japanese Leaners (CEJL[20]) on kogub inglise keelt õppivate jaapanlaste kirjalikku ja suulist keelekasutust. Korpus koosneb neljast allkorpusest, millest üks sisaldab 12–19 aastasete laste ja noorte tekste.

Standard Speaking Test Corpus (SST[21]) sai alguse Jaapanis 1999. a, eesmärgiga koguda 1 miljoni sõna suurune jaapanlastest inglise keele õppijate loomulikku suulist keelekasutust peegeldav korpus. Korpuse valmimise tähtaeg oli aastal 2004. (Ellis, Barkhuizen 2005: 241)

Hiina keele õppija korpus

Hong Kong Corpus of Primary School Chinese (HKCPSC[22]) sisaldab hiina algklassiõpilaste tekste. (http://shs.hku.hk/corpus/main.htm)

Mitmekeelsed laste- ja õppijakeele korpused

Child Language Exchange System (CHILDES[23]) – rahvusvaheline tekstiarhiiv, mis sisaldab suulist kõnet erinevas vanuses kas emakeelt teist keelt või võõrkeelt

[19] http://www.flloc.soton.ac.uk/

[20] http://www.lb.u-tokai.ac.jp/lcorpus/

[21] http://leo.meikai.ac.jp/~tono/sst/index.html

[22] http://www.ncbi.nlm.nih.gov/pubmed/18415738

[23] http://childes.psy.cmu.edu/

kõnelevatelt lastelt. Korpus sisaldab ka kümne 1–3 aasta vanuse eesti lapse spontaanse kõne lindistusi, kokku 154 lindistust. (Vija 2007: 374) Korpus on avalik.

Visual Interactive Syntax Learning (VISL[24]) on omalaadne 25 keelt sisaldav korpus, mille süntaktiliselt märgendatud lausete baasil on loodud elektroonilised õppemängud.

Tabel 5 annab ülevaate mõningate tuntumate õppijakorpuste suurusest.

Tabel 5. Andmed tuntumate õppijakorpuste kohta (Diaz, Fernandez 2006: 87).

korpuse nimi	**korpuse suurus sõnades**	**õppijate emakeel**	**korpuse eesmärk**	**meedia**	**õpitav keel**
CLC	20 000 000	erinevad keeled	kommertsiaalne	kirjalik	inglise
C-LEG	28 000	inglise	akadeemiline	kirjalik	saksa
FALCO	36 000	ei ole määratud	akadeemiline	kirjalik	saksa
FRIDA	200 000	erinevad keeled	akadeemiline	kirjalik	prantsuse
ICLE	2 000 000	erinevad keeled	akadeemiline	kirjalik	inglise
LLC	10 000 000	erinevad keeled	kommertsiaalne	kirjalik	inglise

[24] http://visl.sdu.dk/

4. Korpuspõhised uuringud

Korpuslingvistika võimaldab korraldada peamiselt kahte tüüpi uuringuid: keele struktuuri ja keele kasutamise uuringuid (Biber *et al.* 2006: 269). Korpused iseenesest ei sisalda keele kohta uut informatsiooni, kuid korpusanalüüsis kasutatavad programmid esitavad keele nüansse ja struktuuri viisil, mis võimaldab senisega võrreldes teistsuguseid uuringuid (Hunston 2002: 3). Iga uurimisviisi tuleb suhtuda kui teooriasse, sest see, mida kajastavad näiteandmed, sõltub tõlgendajast.

Korpusuuringuid võib jagada ka **kvantitatiivseteks** ja **kvalitatiivseteks**. Valdav osa korpuse põhjal tehtud analüüse on kvantitatiivsed. (Biber *et al.* 2006: 4–9). Kvalitatiivanalüüsiga on võimalik uurida, kuidas üht või teist sõna (fraasi, grammatilist vormi vm) tekstis kasutatakse ja analüüsida keeles esinevaid seaduspärasusi (Keeffe *et al*. 2007: 2–3).

Korpusuuringute meetoditeks on statistiline analüüs, konkordantsiridade tõlgendamine, kollokatsioonide analüüs ja korpuse annotatsioon. Esimesed meetodid baseeruvad sõnal või sõnavormil, esitledes esinemismustreid ja koosesinemisi. Neist viimane on kõige keerukam meetod, mis eeldab, et sõnad jagatakse süntaktilistesse ja semantilistesse kategooriatesse. (Hunston 2002: 92)

Kõige suurem avastus, milleni korpuste uurimise abil seni on jõutud, seisneb arusaamas, et sõnade sagedusjärjekord tekstis ei ole juhuslik (Hunston 2002). Leksikaalsete ja grammatiliste sõnade seotus põhineb vähemalt kolmel seaduspärasusel:

- grammatilised sõnad on sagedasemad kui leksikaalsed sõnad;
- grammatilised sõnad esinevad suurel määral paradigmades, leksikaalsed sõnad süntagmades;
- grammatilised sõnad võivad samuti kanda sisutähendust nagu leksikaalsed sõnad.

Õppijakorpust koostades peetakse enamasti silmas võimalust uurida keelekasutust, keelelisi mõjutusi ja keele muutumist. Uurimisviisidest on kasutusel võrdlev analüüs ja õppija keelevigade analüüs. Õppijakorpuste põhjal tehtud uurimused on analüüsinud

õpilastekstides kasutatavate keelendite statistikat (tekstitüüpe, autoreid, situatsioone ning kogumismeetodeid varieerides), sõnaliigiti esinevaid vigu, süntaksit, pealausete ja nende juurde kuuluvate osalausete pikkusi, passivi kasutamist, nominalisatsiooni, kollokatsioone, üks kord tekstis esinevaid sõnu, ajaväljendeid, asesõnu, minevikku, eitust, segunemisvigu, vealiike jm. (Nesselhauf 2005, Granger 2007, Biber 2006)

Võrdlev analüüs ehk kontrastiivanalüüs

Kontrastiivanalüüsi tulemus sõltub võrdlusalusest. Õppijatekstide puhul võivad võrdlusaluseks olla kas ekspertgrupi tekstid, mõne teise õppijagrupi tekstid või õppijate jaoks kirjutatud tekstid.

- õppijagrupp ning ekspertgrupp (Hunston 2002: 206).

Õppija keelt võrreldakse emakeelse täiskasvanu keelega. Oluline on arvestada täiskasvanute keele formaalsuse astet, murdetausta, teemavaldkonda jm.

- erinevad õppijagrupid (Hunston 2002: 206).

Õppija keelt võib võrrelda teise õppija keelega vanusest, piirkonnast või teksti loomise situatsioonist lähtuvalt (Granger 2007: 52–53 põhjal).

- õppija keel ja õppija jaoks kirjutatud tekstide keel (Römer 2004: 151).

Uuritakse, kuidas mõjutab õppijat sisendkeel: õpikutes, ajakirjanduses, ilukirjanduses, reklaaminduses, filmides, laulusõnades kasutatav keel, murdeline keelekeskkond vm.

Õppija keelevigade analüüs

Õppijakorpuste peamine uurimismeetod on keelevigade analüüs (ingl *computer-aided error analysis (CEA)*): keelekasutust uuritakse selle grammatilist või semantilist õigsust hinnates.

Vigadeks peetakse (Corder 1981: 5–9; 56–58) selliseid keelendeid, mida õppija ise parandada ei oska (täpsemalt vt peatükk 5). Vigade analüüsi on taunitud kui iganenud viisi õppijakeelele lähenemiseks. See on seotud varasema suhtumisega, et õppija ei tohi vigu teha (Meunier 2007: 27). Praegu valitseva konstruktivistliku arusaama kohaselt rakendab õppija õpitava nähtuse mõistmiseks olemasolevaid mõtteskeeme (ingl

assimilation). Kui tekib viga ja mõtteskeem ei toimi, püüab ta oma arusaama uuendada ja täiustada (ingl *accommodation*). Õpilase uued kogemused peavad senisesse teadmiste süsteemi (ingl *equilibrium*) sobituma, uus tuginema tuntule. Õppija keeles esinevad vead peegeldavad keelesüsteemi, mille õppijad on loonud (teise keele ja võõrkeeleõppes nimetatakse seda *vahekeeleks*) (Corder 1981: 5–13; 57–64). Kui emakeelt või võõrkeelt alles õpitakse, ei saagi keelekasutus olla päris üldkeele reeglitele vastav.

Vead näitavad keeleuurijale, kuidas õppimine toimub – kuidas õppija katsetab õpitava keele kohta käivaid hüpoteese. Veaanalüüsi uurimistulemused on olulised ka õppijate endi jaoks, kuna analüüsist tehtavate järelduste põhjal on võimalik õppematerjale õppija vajadustele vastavalt parandada (Corder 1981: 77) ning ainekavade sisu korrigeerida.

Teine põhjus õppija vigade uurimiseks tuleneb laste ja noorte vastuvõtlikkusest kõigele uuele. Õppijate keel peegeldab väga kiiresti ajakirjandus- ja internetikeele mõjul üldkeeles toimuvaid muutusi. Õpilaste kirjandite põhjal on uuritud kasutatavat sõnavara: E. Paomees on uurinud suulise kõne ilminguid eesti gümnasistide lõpukirjandites ning leidnud, et 2/3 kirjutajatest kasutab oma töös suulise kõne sõnavara. Enimkasutatud suulise kõne sõnad olid *asi, värk, sina, peale, tahaplaanile, läbi, ennem, vahest, õieti, jama*. (Paomees 2003: 43–47). K. Lepajõe (2006: 141–153) on analüüsinud õppijate registri- ja stiilitunnet ning kirjandite *sina*-deiksist. E. Jüssi on õppijate keelekasutuses täheldanud laadivahelduse taandumist (Jüssi 2005).

Külli Habicht on 2005.–2006. aasta gümnaasiumilõpetajate eksamikirjandeid analüüsides käsitlenud tüüpilisi morfosüntaktilistest normingutest hälbivaid tarvitusjuhte ning põhjendanud neid universaalse ja keelesisese loomulikkustaotlusega. Esile on toodud ainsuse ja mitmuse osastava käände vormide markeerimine, lühikese ülivõrde vormide kasutamine, astmevaheldusmallide lihtsustumine, verbiühendite kokkukirjutamise tendents, mineviku kesksõnade adjektiivistumine, kaassõna ja sellele eelneva noomeni kokkukirjutamine. (Habicht 2006)

K.Habichti ja P. Penjami 2006. aastal korraldatud sünkrooniline uuring keskendus kaassõna kokku- ja lahkukirjutamise juhtudele. 205 õpilase ja 194 üliõpilase seas korraldatud küsitlus näitas, et kaassõna kirjutati eelnenud nimi- või asesõnga sõnaga

kokku 21,4% juhtudest. Sagedamini (üle 20% juhtudest) kokku kirjutatud kaassõnafraasid (*metsa vahel, pea kohal, varga kombel, mehe moel, minu moel, sinu kombel, selle kõrval*) on tõenäoliselt grammatiseerumas, mistõttu neile omistatavat tähendust üritatakse väljendada ka kirjapildis (Habicht, Penjam 2006: 51–62; Jürine 2007: 30).

4.1. Märgendamata korpus

Märgendamata korpust on võimalik uurida sõnaloendite (sagedusridade) või konkordantsiridade abil. (Bowker, Pearson 2002: 13; Hunston 2002: 20–22) Korpusepõhiste uuringute tegemiseks on vaja vastavaid programme.

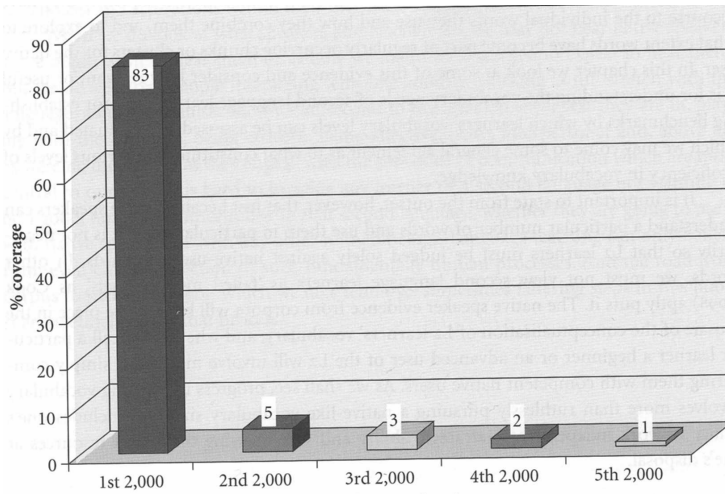
4.1.1. Sagedusloend

Sagedusloend annab ülevaate uuritava(te)s teksti(de)s või tekstitüüpides esinevate sõnade tegeliku esinemissageduse kohta ja võimaldab teha statistilisi uurimusi: palju on tekstis sõnu kokku, mitu korda mingi tähejärjend (*string*) kordub. Õppijakeele korpuse statistiline sõnasagedusanalüüs peegeldab teatud sõnade või grammatiliste vormide üle- või alakasutust õppija keeles (vt lisa 5). (Barlow 2005: 335; Tribble, Johns 1997: 23–24.) Sõnasagedusloendi abil saab märksõnu sorteerida kas nende esinemise järjekorras, tähestiku või esinemissageduse järgi (Bowker, Pearson 2002: 13).

Lemmatiseerimata ehk algvormidele taandamata teksti puhul võiks kasutada tähestikulises järjekorras esitatud loetelu, et sarnase tähejärjestusega sõnad koonduksid ning annaksid sel moel sõnasagedusest parema ülevaate. Sagedusloend ei suuda eristada vormihomonüümiat ega leksikaalset homonüümiat, mistõttu sobib paremini aglutinatiivse keele jaoks, kus tüvemuutusi pole. Sagedusloendi põhjal on võimalik analüüsida õpilaste sünonüümikasutamisoskust (Tribble 1997: 24). Sõnu on võimalik järjestada ka sufiksi järgi. (Bowker, Pearson 2002: 116–117)

Sagedusloend on oluline infoallikas keeleõppematerjalide koostajate jaoks, kuna kehtib seaduspära, et paar tuhat sagedamat lemmat (*lemma* – sõna algvorm)

moodustavad uuritavast tekstist alati u 80% (vt joonis 3). Järgmised paartuhat lemmat katavad teksti sõnakasutusest juba üsna tühise osa.



Joonis 3. Sagedusloendi esimesed paar tuhat lemmat katavad keelekasutusest olulise osa. Joonis on tehtud 10 miljoni sõna suuruse korpuse (Cambridge International Corpus ja CANCODE) sõnade sagedus- ja kasutusuuringu põhjal (O´Keefee 2007: 32).

Kirjakeele sagedussõnastikus esitatud lemmade tabel näitab sageduse järgi reastatud lemmade kumulatiivset osakaalu teksti katmisel (vt tabel 6). Selgub, et u 1 miljoni sõna suuruse korpuse 250 kõige sagedasemat sõna katavad üle poole tekstist ja 10 000 kõige sagedasemat sõna u 90% tekstist. Üle poole lemmadest esines tekstis ainult üks kord.

Tabel 6. Lemmade kumulatiivne osakaal teksti katmisel. Numbrid on ümardatud (Kaalep, Muischnek 2002:14**).**

Esimesed ..... sõna	Mitu % tekstist nad katavad
10	19,3
20	24,6
50	33,1
100	40,7
250	51,3
500	60,2
1000	69,0
1500	74,0
2000	77,2
3000	81,5
5000	86,0
10 000	90,3

Kirjaliku keele omapärana on ilmnenud, et igapäevases kodukeeles tavalised sõnad (*laud, tool* vms) ei ole korpuses sagedamini esinevate sõnade seas. Selle põhjuseks võib

pidada nö kodutekstide vähest esindatust korpuses. Seega ei garanteeri sagedus mingis tekstiklassis sõna tavalisust (Kaalep, Muischnek 2002: 9).

Kuna sõnad esinevad tekstis vastavalt teksti teemale, siis mistahes tekstide alusel leitud sõnasagedus esitab tavalisusest moonutatud pildi. Lisaks sagedusele tuleb arvestada sõna levikut erinevates tekstides. Sõna, mis esineb paljudes tekstides on sagedam kui sõna, mis esineb ühes tekstis väga palju kordi. (Kaalep, Muischnek 2002: 9)

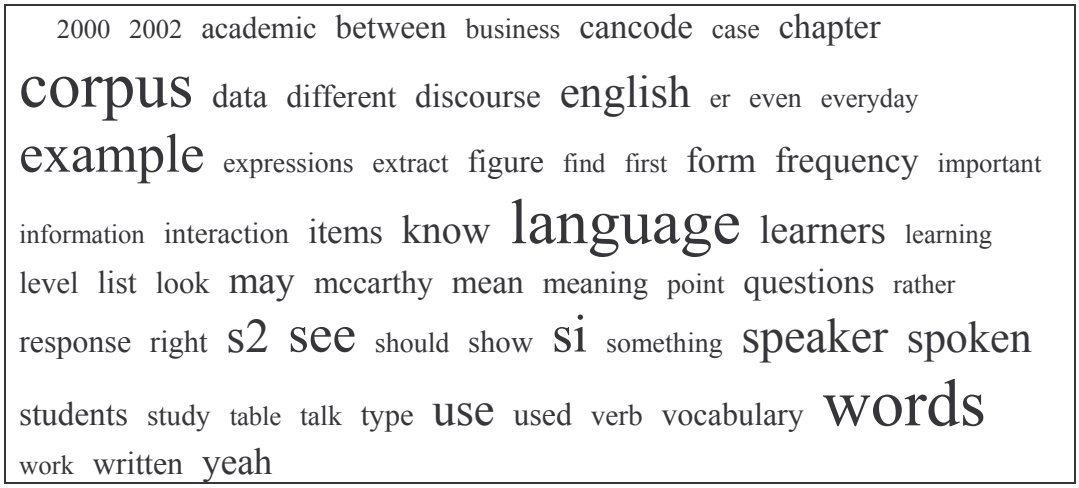
Keele õppimise seisukohalt on lisaks sõnade esinemissagedusele oluline ka nende tavalisus, keerukus, otstarbekohasus, õppimise hõlpsus, õppija eesmärgid, vajadused jm (Nesselhauf 2005: 255; Conrad 2007: 63). Emakeeleõppes pakuvad sagedate esinemiste kõrval huvi ka harvaesinevad sõnad, kuna õppijad võivad teadlikult vältida sõnu ning väljendeid, mis on vähemkasutatavad, keerulise tähendusega, rasked hääldada, keerulised kirjutada vm ning eelistada sõnu, mis on lihtsad ja tavalised (Sökmen 1997: 239–240 viidatud Granger 2007: 62 järgi). Huvipakkuv on ka grammatiliste konstruktsioonide sagedusanalüüs, sest kuigi inimestel on teatud intuitsioon selle kohta, missugused sõnad sagedamini esinevad, puudub intuitsioon grammatiliste konstruktsioonide esinemissageduse kohta .

Sisu- ehk võtmesõnade statistiline analüüs (ingl *key word analysis*)

Sisusõnad on sõnad, mis esinevad konkreetses tekstis tavalisest sagedamini, või mille sagedus konkreetses tekstis on ebatavaliselt kõrge, kuigi need ei ole teksti sagedaimad sõnad. (Keeffe *et al.* 2007: 12, Hunston 2002: 68). Sisusõnade analüüsiga võib otsida mitmest sõnast koosnevaid fraase, võrrelda korpuse erinevate osade sisusõnade loendeid omavahel või mõne teise korpuse loendiga (näiteks erinevas vanuses õpilaste töid omavahel või õpilaste töid võrreldes üldkorpusega)(vt ka emakeeleõppijate eksperimentaalkorpuse statistilist analüüsi pt 6.2.3.).

Võrdlevate loeteludega saab esile tuua korpustes kasutatud keele statistilised erinevused (näiteks teatud vormide, väljendite üle- ja alakasutus) (Granger 2007: 55). Sisusõnade analüüsi abil saab erialatekstidest selekteerida põhiterminid, kõnest parasiitväljendid, uurimuslikust tekstist uurimuslikule tekstile kohased väljendid jne. Märksõnaloendid on olulised stilistikas ja võrdlevas tekstianalüüsis. Leksikaalsete üksuste kõrval tulevad sisusõnadena esile grammatilised sõnad.

Sisusõnu on võimalik esitleda sagedusloendina selliselt, et sagedamini esinevad sõnad esitatakse nimekirjas eespool (Bowker, Pearson 2002: 31) või visualiseeritult (vt joonis 4). Visuaalne esitlusviis on atraktiivne ning samas informatiivne. Mida rohkem kordi sõna tekstis esineb, seda suurema kirjaga see esitatakse. Elektroonilisel kujul esitatud visualiseeritud pildil on märksõnad nö aktiivsed, st kursoriga vastava sõna peale liikudes on näha sõna esinemiste arv tekstis. Näiteks võib joonisel 4 esitatud märksõnade põhjal väita, et esitletav ingliskeelne raamat käsitleb teise keele õppija suulist keelekasutust korpusepõhiselt.



Joonis 4. Tekstis esinevate märksõnade visualiseerimine[25].

Visualiseeritud märksõnaesitlust võib kasutada õppijate eeskujukorpuses erialatekstide sisu tutvustamiseks. Programm võib esitleda ühe sõnaloendi mitme faili kohta korraga või iga teksti kohta eraldi.

[25]Vaata *www.amazon.ca*

4.1.2. Konkordants

(Kwic-Format, Keyword in Context, KIWIC, KWIC-format)

Konkordantsid ehk märksõnaotsingud analüüsivad sõna nende esinemise kontekstis. Konkordantside abil on keelekasutust uuritud ja keeles esinevaid mustreid esitletud juba üle 25 aasta (Johns 2007: 4; Hunston 2002: 42). Konkordantsiotsinguks on spetsiaalsed programmid, mis esitavad sõna (või fraasi) koos seda ümbritsevate sõnadega – nö sõna kontekstis. Märksõnaotsinguga saab korpusest kiiresti leida kõik otsitava sõna esinemisjuhud (Bowker, Pearson 2002: 13) ja jälgida otsitava esinemise regulaarsust, mis kipub jääma märkamatuks, kui sama fraas või sõna on teksti sees.

Konkordantsiridadega saab otsida lemmat, sõna või fraasi. Ingliskeelses kirjanduses tutvustatakse märksõnaotsingut, mis esitab otsingutulemuse sel moel, et otsitav sõna jääb veergu lehekülje keskele. Ümbritsev kontekst on määratletud esitatavate tähemärkide arvuga (näiteks 80). Sel moel esitatud kontekstis on laused sageli poolikud, alates isegi poolelt sõnalt (vt joonis 5). Seaduspärasuste leidmiseks ei ole vaja teksti lugeda lineaarselt nagu Euroopa kultuuriruumis tavaliselt kombeks. Konkordantsiridades tuleb jälgida vaid esiletõstetud sõna ning sellega vahetult või kaudselt seotud sõnu.

1	o the locations of their sources. It is **interesting** to speculate what the source
2	d produce a top quark. The rest, though **interesting** for a host of other projects
3	silicates. Carbonates Are Key The most **interesting** aspect of ALH84001 are the c
4	ty-three days." JUMPING FISH--"The most **interesting** examples of amphibious fishe
5	the mechanisms of its action will prove **interesting** to deduce. Some studies have
6	cally, but it would still be undeniably **interesting** and valuable to know how a p
7	ids. NANCY COLLINS (on camera) It`s **interesting** because there are so many
8	several of the bidders. They had some **interesting** reasons for why they were he
9	s great crime prevention, and it's very **interesting**. ROBERT KRULWICH (intervie
10	may be part of it. However, there is an **interesting** development in Hollywood, wh
11	on. "I've got it. You don't." With one **interesting** exception. If a guy is in a
12	DONALDSON By the way, here's something **interesting** -- James Patterson says
13	born again. So it was a really, really **interesting** thing. You know in the Pente
14	OKS Well, I think that one of the other **interesting** objects in the sale that the

Joonis 5. Ingliskeelses kirjanduses tutvustatud konkordantsiotsingu tulemuse näide[26].

[26] Sõna *interesting* konkordantsiotsing 17 miljoni sõna suuruses Nagaoka Koseni korpuses. *www.nuis.ac.jp*

TÜ keeletehnoloogide koostatud korpuse kasutajaliides esitab märksõna terviklauses. Märksõna on küll mõneti halvemini märgatav, kuid ümbritsev kontekst on terviklikum ning sõna tähendus tuleb paremini esile (vt joonis 6). Kui uurija vajab laiemat konteksti, võib programm selle esitada ka lõikude kaupa.

AJAE1980\tat0521 Küüditatute puhul **huvitav**ad arvud vaevalt küll laiemat lugejaskonda, huvi keskpunktis seisab inimene.

AJAE1980\tat0525 Minu arvates peaks maleajakirjanik leidma turniiril **huvitav**at informatsiooni ka veretutest viikidest.

AJAE1980\tat0525 Nende endi sõnul pakkus see töö suurt rahuldust, sest on väga **huvitav** näha, kui tugevaid emotsioone maletajad siiski üle elavad.

AJAE1980\tat0765 Stiililiselt sobisid küll lõputult üht ja sama liigutust tegev mees ja Rothenbergi minimalistlik fraasikordus kokku, kuid elava muusika tegid **huvitav**aks minimaalsed muutused monotoonses ahelas, surnud pilt filmilindil, aga ammendas end kiiresti, hoolimata näitaja pingutustest seda erinevalt fraseerides **huvitav**amaks teha.

AJAE1980\tat0769 Vahvad laulud olid (ja on), veidi sentimentaalsed, naiivsedki, **huvitav**a lahtise, ootamatu huumoriga ning alati täis elujõudu: Kui on saatus sul karm, petnud õnn sind või arm..., Viin, su terviseks joon..., Täna on sinine esmaspäev..., Torisev mesilane Sebedeus mõtles..., jne.

AJAE1980\tat0602 Soome teatrimehed, kellel on tavaks igal aastal käia maailma teatrikeskustes kõike uut ja **huvitav**at vaatamas, olid lummatud.

Joonis 6. TÜ keeletehnoloogide otsinguprogrammi päringutulemuse näide[27].

Mõne otsingu puhul on vaja ridu toimetada või sorteerida, et otsinguga esitatud suurest andmehulgast eraldada otsitavad fraasid ja kõik korduv või tüüpiline fikseerida. Kui andmehulk on tõesti väga mahukas, on hüpoteeside püstitamiseks ja nende kontrollimiseks otstarbekas materjali väiksemate osade kaupa analüüsida (Hunston 2002: 9, 65).

Konkordantsiotsinguga saadud info põhjal on võimalik koostada sõna leksikogrammatiline ülevaade, mis esitab kontekstis esineva sõna

1. kollokatsioonid: koosesinemised, mis on analüüsitava sõna jaoks kõige tavalisemad ja statistiliselt tähtsamad (mitte juhuslik esinemus);
2. idioomid: kas sõna on mõne idioomi koostises;

[27] vaata *www.cl.ut.ee/korpused*

3. süntaktilised kitsendused: kas sõna esineb lauses teatud positsioonil, missugused on sõna süntaktilised piirangud;

4. semantilised piirangud;

5. semantiline prosoodia – kas sõnal on semantilisi kollokatsioone, kas sõna on pigem positiivse või negatiivse värvinguga jmt.

6. muud korduvad esinemismustrid. (Keeffe *et al.* 2007: 14–15)

4.1.3. Kollokatsioon

Kadri Muischnek (Muischnek 2006b: 17) on kollokatsioone defineerinud kui sõnade või sõnavormide esinemust üksteise naabruses sagedamini, kui võiks eeldada nende sõnade endi sageduse põhjal. Grammatilisi malle võib leida alates lihtsatest leksikaalgrammatilistest kombinatsioonidest (näiteks verb ja tema laiendid) kuni komplekssete süntaktiliste mallideni. (Muischnek 2006b: 17) Kollokatsioon võib olla semantiliselt põhjendatud – näiteks esineb sõna *mänguasi* sagedamini koos sõnaga *laps* kui *mees*, või mittepõhjendatud – näiteks sõna *kange* esineb koos sõnaga *tee* (*kange tee*) palju sagedamini kui sõnaga *auto* (*? kange auto*). (Hunston 2002: 12–13)

Kollokatsioonide sagedusanalüüsiga eristuvad enamasti grammatilised kollokatsioonid (vt joonis 7).

French	German	Czech	Times
of the	of the	of the	of the
in the	in the	in the	in the
to the	to be	it is	to the
to be	to the	to be	on the
it is	it is	do not	for the
of a	on the	to the	to be
is a	have to	It is	and the
on the	of a	is the	at the
will be	in a	is a	that the
is the	and the	is not	of a

Joonis 7. Õppijatekstide sagedaimad kollokatsioonid prantsuse, saksa ja tšehhi õpilastekstides võrreldunа ajalehe Times sagedaimate kollokatsioonidega. (Ellis, Barkhuizen 2005: 353).

Inglise keeles on kollokatsioone võimalik otsida spetsiaalse programmi abil, mis näitab otsitava sõna ümbruses sagedamini esinenud sõnade hulka (Bowker, Pearson

2002: 126). S. Hunston kirjeldab, kuidas kollokatsioonide arvutamise programm arvutab kokku märksõna (*node word*) ümbruses (näiteks nelja sõna kaugusel paremal ja nelja sõna kaugusel vasakul) asuvate sõnade esinemuse ja esitab märksõnaga koos esinenud sõnade sagedusrea. Sagedaimad on tavaliselt grammatilised sõnad. (Hunston 2002: 68–75) Automaatse kollokatsioonide analüüsi ebaõnnestumist on kirjeldanud Nadja Nesselhauf, kelle väitel jäävad paljud kombinatsioonid esitamata, kuna kokkukuuluvad sõnad on üksteisest veidi kaugemal. Seevastu esitatakse automaatse analüüsiga hulgaliselt kasutuseks sobimatut materjali, mille sorteerimine on eriti aeganõudev. (Nesselhauf 2005: 43)

N. Nesselhaufi (2005: 11–12) väitel kasutatakse mõistet *kollokatsioon* lingvistikas ja keeleõppes erinevas tähenduses ning mõistele võib anda kaks selgitust. Mõlemal juhul märgib *kollokatsioon* midagi ülesehituslikku, konstruktsioonilist. Esimesel juhul on tegemist sagedusel põhineva koosesinemisega: sõnade koosesinemine sagedamini, kui võiks eeldada nende sõnade tavapärase juhusliku koosesinemise põhjal. Teisel juhul on koosesinemine nö fraseoloogiline: sõnade kombinatsioon, mis on teatud määral fikseeritud. (Nesselhauf 2005: 11–12). Sõnad võivad koos esineda mitmetes kombinatsioonides, kuid ainult üks neist on kinnistunud mingi tähenduse väljendamiseks.

Fraseoloogilisi kollokatsioone võib teistest nimisõna ja verbi ühenditest eristamiseks jagada (Nesselhauf 2005: 32–33) kolme suurde rühma:

1) vabad ühendid (*loeb lehte*), kus nii verb kui nimisõna on vabalt ja piiranguteta muudetavad;

2) kollokatsioonid (*kehitas õlgu, teeb otsuse*), milles nimisõna on vahetatav ilma piiranguteta, kuid verbil säilib sama tähendus vaid teatud nimisõnaga;

3) idioomid (*astus ämbrisse*), mille puhul ei ole ei verb ega nimisõna vabalt vahetatavad, kuna sel juhul ei säili väljendi tähendus.

4.2. Korpuse märgendamine

Sageli on vaja, et elektroonilisel kujul olevad tekstid sisaldaksid eksplitsiitsel kujul veel mingit muud informatsiooni peale selle, mis originaaltekstides esialgselt olemas on (Kaalep 1999: 13; Bowker, Pearson 2002: 83). Korpusesse info lisamist nimetataksegi märgendamiseks (Muischnek *et al.* 2003: 12). Tekstidele võib lisada teavet näiteks

ülesehituse (peatükid, pealkirjad, lõigud, laused jne), morfoloogilise ja süntaktilise analüüsi tulemuste vm kohta. Teksti lingvistilist märgendamist nimetatakse annotatsiooniks (*annotation,* ka *tagging, mark-up*) ning lingvistiliselt märgendatud teksti annoteeritud tekstiks (Bowker, Pearson 2002: 83–84).

Korpuse märgendus peab olema ühtlane kogu märgendatava korpuse osa ulatuses. Näiteks ei tohi lausete märgendamisel tõlgendada otsekõne lauset ja saatelauset kord kaheks omaette lauseks, kord üheks. Kuna märgendamine on väga töömahukas tegevus, siis on mõtet märgendada nii vähe kui võimalik, ehkki nii palju kui vajalik. (Kaalep: 1999: 19–20)

Nö puhta teksti poolt oli Sinclair (1995: 23), kes väitis, et igasugune märgendamine on teksti tõlgendamine (mööndes, et teatud ulatuses ja kindla uurimise eesmärgil tuleb seda teha). Kuna märgendused on tekstis näha (vt joonis 8), võivad hakata need uurijat segama ning seepärast on vaja, et ka nö puhas tekst oleks kättesaadav. (Bowker, Pearson 2002: 86)

L'héritage du passé est très <G><GEN><ADJ> #fort$ forte </ADJ></GEN></G> et le sexisme est toujours présent. Beaucoup de gens pensent que la femme est un être pas très intelligent, qui bavarde beaucoup et qui ne sait pas garder le moindre <F><DIA><NOM> #secret$ secrèt </NOM></DIA></F>. Ces gens <G><NBR><VSC> #pensent$ pense </VSC></NBR></G> aussi que les femmes ne sont pas aptes à prendre des responsabilités.

Joonis 8. Õppijakorpuse FRIDA veamärgendustega tekst (Granger 2003: 470).

Teksti märgendamiseks on kolm viisi: automaatne, poolautomaatne ja käsitsi märgendamine. Esimeste eeliseks on kiirus, viimase eeliseks põhjalikkus. Suuremate korpuste puhul on eelistatud standardse tarkvara abil tehtav automaatne või poolautomaatne märgendamine. Näiteks sõnaliigi määramiseks kasutab programm oma andmebaasi (sõnastikku) ehk võrdleb tekstist leitud sõna sõnastiku omaga ja otsustab siis, mis sõnaliigiga tegu on. Kui sõnastikus sellist sõna ei ole, oletab programm talle kirjeldatud reeglistiku alusel, kuhu sõna võiks kuuluda. Kui võimalusi on mitu, esitatakse kõik võimalused. Kui programm on sõnadele märgendus(ed) lisanud, vaatab uurija need üle ja teeb oma parandused. (Bowker, Pearson 2002: 85)

Automaatne märgendus on kõige kiirem, kuid ka kõige ebatäpsem. Käsitsi märgendamine on aja- ja tööjõukulukas, kuid järjekindluse ja selgete reeglite järgimisel

annab hea tulemuse. Märgendamise alused tuleb lahti kirjutada ning lisada näited. (Ellis, Barkhuizen 2005: 267) Kui märgendamise alus töö käigus muutub, tuleb tehtud töö uuesti teha.

Sõltumata sellest, mida tahetakse märgendada, on mõistlik alustada /.../sellisest, mille lisamine on võimalikult automatiseeritav ja üheselt mõistetav. Näiteks soovitab H.-J. Kaalep struktuuri märgendamise puhul alustada osadest ja peatükkidest, seejärel märgendada lõigud, loendid, luuletused /.../ ja siis alles laused. (Kaalep 1999: 14) Märgendada võib ka ainult ühte osa korpusest.

Geoffrey Leech (1997, viidatud Lemnitzer 2007: 96 kaudu) on sõnastanud viis märgendamise maksiimi ehk põhimõtet:

- Märgendus peab olema sisestatud selliselt, et puhas märgenditeta tekst oleks taastatav.
- Märgendeid peab olema võimalik analüüsida tekstist sõltumatult.
- Märgendamise alused tuleb dokumenteerida. Dokumentatsioon peab olema märgendajatele ja korpuse kasutajatele kättesaadav.
    - A) vajadusel peavad nii märgendajad kui ka märgendamise viis (automaatne, käsitsi vm) olema tuvastatavad
    - B) ükski kasutatav märgendamise süsteem ei ole ainuõige.
- Märgendamise alused peaksid peavad tuginema teoreetiliselt neutraalsele analüüsile.
- Ühtegi märgendamise süsteemi ei saa *a priori* standardiks kuulutada.

Teksti eeluuring – kvantitatiivanalüüs esitab sõnade ja lausete pikkused, tähtede arvu tekstis või sõnades keskmiselt, lausete arvu tekstis, sõnavormide sagedusloetelu, sagedamini esinevate sõnade loetelu vms (Bowker, Pearson 2002: 109).

Lausestamine (ingl *tagging*). Igasugune märgendamine algab teksti jagamisest lõikudeks ja lauseteks. Lausepiiride leidmine võib tunduda triviaalse ülesandena, kuid seda ta siiski pole. Näiteks punkt numbri, lühendi või initsiaali taga võib, aga ei pruugi tähistada lause lõppu. Samuti võivad lauselõpupunktile järgneda sulud, jutumägid või veel mingid muud sümbolid, mistõttu lause lõpp tuleb alles pärast punkti. (Kaalep, Vaino 2000: 90) Lausestamine on tehtav täisautomaatselt (Muischnek *et al.* 2003: 13).

Tehniline ja ortograafiline märgendamine – märgenditega eraldatakse tekstiosad – peatükid, lõigud, laused, pealkirjad, allkirjad, pärisnimed, lühendid, numbrid, kuupäevad. Teksti selguse huvides asendatakse kõik tekstis esinevad jutumärgid ühetüübilistega (Muischnek 1998: 9).

Lemmatiseerimine on sõnavormide viimine algvormile. Eesti keeles on statistilise ühestaja ning morfoloogilise analüsaatori kasutamisel võimalik lemmasid automaatselt leida, kuid ei ole võimalik eristada sõnaliike ega homonüüme (vt Kaalep, Vaino 2000: 92–98). Näiteks koonduvad omadussõna *noor* ning nimisõna *noor* ühte lemmasse.

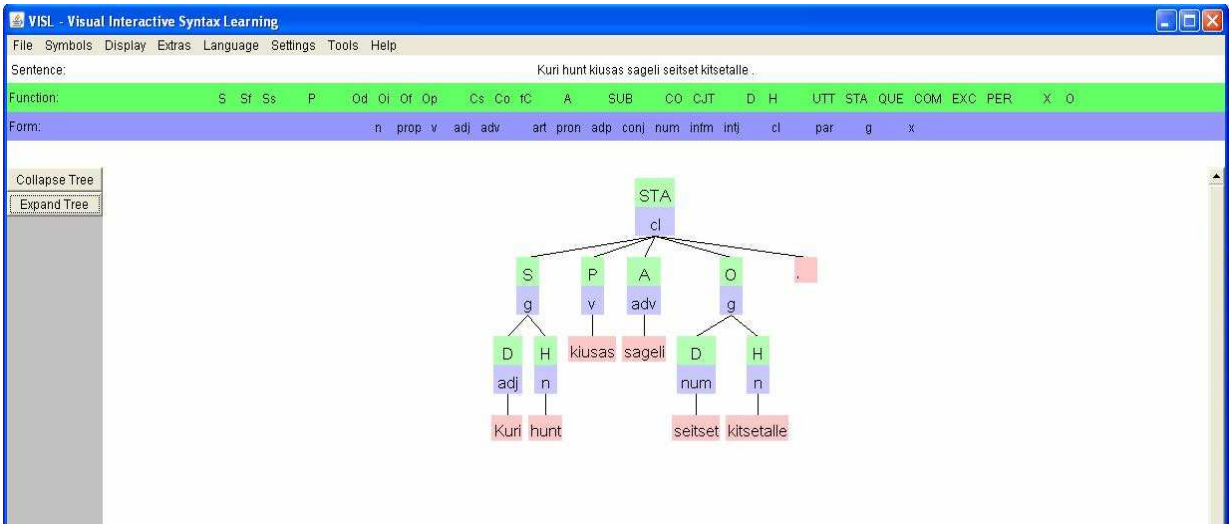
Morfoloogiline märgendamine on üks levinumaid märgendamise viise. Sõnale lisatakse info tema algvormi, sõnaliigilise kuuluvuse ja grammatiliste kategooriate kohta. Inglise keele kui väga vähese muutemorfoloogiaga keele puhul kasutatakse sageli morfoloogilise märgendamise lihtsustatud varianti ehk ainult sõnaliigilise kuuluvuse märgendamist (ingl *part-of-speech tagging, POS-tagging*). Morfoloogiline märgendamine võimaldab uurida teksti sõnaliikide kaupa (Hunston 2002: 81), teostada morfoloogilist analüüsi ja ühestamist. Eestikeelset teksti saab morfoloogiliselt märgendada peaaegu täisautomaatselt (Muischnek *et al*. 2003: 13) (vt joonis 9).

kitsenduste grammatika märgendussüsteem[28]	**eesti keele morfanalüsaatori analüüs**
Pühapäeva püha_päev+0 //_S_ com sg gen #cap // õhtul õhtu+l //_S_ com sg ad // esinesid esine+sid //_V_ main indic impf ps3 pl ps af #FinV #Intr //	Pühapäeva püha_päev+0 //_S_ adt, sg p, // püha_päev+0 //_S_ sg g, // õhtul õhtu+l //_S_ sg ad, // esinesid esine+sid //_V_ sid, //

Joonis 9. Morfosüntaktilise märgendamise näited.

[28] http://www.cs.ut.ee/~kaili/Korpus/proov.txt

Süntaktiline märgendamine (*parsing*) on teine kasutatavaim märgendamise viis – tekstile lisatakse süntaktiline info, märkides iga sõna juurde, millist funktsiooni see lauses täidab. Süntaktilist infot võib esitada ka lause fraasistruktuuri ehk puustruktuurina. Süntaktilist analüüsi tehakse tänapäeval pool-automaatselt. Õpiprogrammis VISL on eestikeelsed laused esitatud puustruktuurina (vt joonis 10).



Joonis 10. Õpiprogrammi VISL lause puustruktuurina.

Semantiline märgendamine võib olla üldjoontes kahte tüüpi: sõnadele lisatakse info mingisse semantilisse klassi kuulumise kohta ehk märgendatakse teksti semantilised suhted. Semantilise märgendamise all võib mõelda ka sõnatähenduste ühestamist, semantiliskategoriaalse kuuluvuse märgendamist (Muischnek 1998: 9, Muischnek *et al.* 2003: 14).

***Ad hoc* märgendamine** – kui uuringu eesmärk nõuab millegi poolest erilist märgendamist (näiteks on vaja tähistada tekstis kõik *last* tähistavad väljendid), siis tuleb seda teha. Taoliselt märgendatud tekst võib anda huvitavaid tulemusi. (Hunston 2002: 94) Tähistada võib ükskõik millist keele aspekti või lingvistilist struktuuri (Ellis, Barkhuizen 2005: 241).

T-üksuste (ingl *T-unit***) kaupa märgendamine.** T-üksus on pealause koos talle alistatud osalausetega. Kaks rinnastatud lauset on kaks T-üksust. T-üksuse pikkust analüüsitakse kui süntaktilise keerukuse näitajat. Kuna automaatne analüüs ei ole võimalik, tuleb T-üksused sisestada käsitsi. (Biber *et al.* 178–180)

Kommunikatsioonistrateegia märgendamine – Burdine on tähistanud erinevad kommunikatsioonistrateegiad ning analüüsinud erinevate strateegiate kasutamise sagedust. (Ellis, Barkhuizen 2005: 341)

Vigade märgendamine ja sagedamate vigade analüüs pakub õpilastekstide puhul tõenäoliselt suurimat huvi, kuna võimaldab esile tuua õppijate keelekasutuse erinevuse üldkeelest ning on väärtuslikuks andmebaasiks efektiivsete õppematerjalide koostamisel. Õppijatekstides esinevate vigade märgendamisest tuleb pikemalt juttu 5. peatükis.

5. Õppijatekstide traditsiooniline uurimismeetod – veaanalüüs

1970. aastatel väga populaarne õppijakeele uurimisviis – vigade analüüs – oli vahepeal ebasoosingus kui häguste piiridega, ebausaldatav ja ebateaduslik viis õppija keelele lähenemiseks. Rakenduslingvistika osana on veaanalüüs automatiseeritav ning läbipaistvam ja sellele on jälle suuremat tähelepanu osutatud. Veaanalüüs on muutunud oluliseks eeltööks õppija vajadustele vastavate õppematerjalide loomisel.

Õpilase keelekasutus erineb emakeelt kõneleva täiskasvanu keelekasutusest nii sõnade ja fraaside poolest kui ka struktuuride sageduselt, kuid keeleõppijate puhul on tähelepanuväärne emakeele õppimise ja teise keele õppimise sarnasus – mõlemad kasutavad nn vahekeele etappi, mil õppija keelekasutus ei vasta õpitavat keelt kasutava emakeelse täiskasvanu keelekasutusele (Corder 1981: 7–8). Õppijakeeles esinevate vigade põhjal saab hinnata, missugused valdkonnad on õppijal veel omandamata. Keele uurija saab õppija vigu analüüsides aimu õppimise protsessi kohta ning õppija saab tagasisidest selle kohta, kas tema ettekujutus õpitava keele kohta on õige (P. Corder 1981: 7–12).

5.1. Viga

Veaanalüüsi tegemiseks tuleb esmalt defineerida, mida peetakse veaks. Keelevea määramisel eristatakse kahte alust

1) vigaseks peetakse vormi, mis erineb üldkeele normist (ingl *grammaticality*). Emakeelse õppija vigade hindamisel on kõige õigem võtta aluseks kehtiv kirjakeelenorm ehk hinnata grammatilisi vigu – üldkehtiva keelenormi reeglite rikkumist. Grammatilise vormi hindamiseks ei ole vaja ulatuslikku konteksti, sest grammatiliselt vigaseks peetakse vormi, mis sellisel kujul üldse kunagi keeles esineda ei saa (James 1998: 65).

2) vigaseks peetakse vormi, mis erineb emakeelekõneleja keelekasutusest (ingl *acceptability*) (Ellis, Barkhuizen 2005: 56). Aktsepteeritava keelekasutuse hindamine tekitab palju probleeme, kuna keel on variatiivne ja keele sees kasutatakse erinevaid

registreid. Kuna viga on sotsiolingvistiline nähtus, saab normingust kõrvalekaldeid, mida emakeelekõneleja ei kasuta, hinnata veana või mitteveana nende kasutamise kontekstist lähtuvalt (Pajusalu 2004: 218). Ka sama registri sees on mitmeid väljendusviise ning ühes kontekstis veana tunduv keelend ei pruugi teistsuguses kontekstis vigane olla (James 1998: 68; Nesselhauf 2005: 38–39).

Näiteks inglisekeelne lause *Pele wore a green dress* sobib karnevalirongkäigus marssija kirjeldamiseks, kuid brasiillaste jalgpallimängijat kommenteerides ei sobi kasutada sõna *shirt* asemel *dress* ning tegemist on sõnavalikuveaga. (James 1998: 67)

R. Ellis (Ellis, Barkhuizen 2005: 56) on teise või võõrkeeleõppijate vigade määramisel lähtunud Lennoni definitsioonist: "Viga on lingvistiline vorm või kombinatsioon, mida emakeelekõneleja samas kontekstis ja samades tingimustes tõenäoliselt kasutanud ei oleks" (Lennon 1991, viidatud Ellis, Barkhuizen 2005: 56 kaudu). R. Ellis kirjeldab analüüsiprotsessi, mille võrreldavateks poolteks on teise keele õppija kirjutatu ning lause, mida samas situatsioonis oleks kirjutanud emakeelekõneleja. Emakeelekõnelejal palutakse muuta õppija tekstis kõik vigased konstruktsioonid, kuid jätta puutumata need, milles viga ei ole. Parandusi tuleb teha nii vähe kui võimalik, st säilitada algset teksti maksimaalselt. Seejärel võrreldakse, mille poolest parandaja lause erineb õppija lausest. (Ellis, Barkhuizen 2005: 56–58)

Kirjeldatud analüüsi puhul ei saa parandaja kunagi kindel olla, et suutis rekonstrueerida mõtte, mida kirjutaja tegelikult öelda või rõhutada tahtis (Corder 1981: 38).

Aktsepteeritavuse tõlgendamine on subjektiivne ja sõltub uurijast. Täiskasvanud keelekasutajad ei pruugi vigaseks või korrektseks pidada samu konstruktsioone. Keelendi tajumine veana erineb kas või sel tasemel, kas registreeritakse ainult õigekeelsusvigu (ingl *absolute errors*) või ka neid keelendeid, mis pole lausa valed, kuid mida ei sovitata kasutada (ingl *dispreferred*). (Ellis, Barkhuizen 2005: 59)

Stilistiliste vigade analüüsimisel jääb keelekasutuse aktsepteeritavus siiski peamiseks otsustamise aluseks.

Ka grammatilisus on üks teguritest, mida keelekasutaja aktsepteeritavuse hindamisel arvesse võtab (James 1998: 67). J. Lyons (1977 viidatud James 1998: 69 kaudu) on esitanud neli grammatikalisuse ning aktsepteeritavuse kombinatsiooni.

Valdav osa lauseid on grammatiliselt korrektsed ning ka aktsepteeritavad

(G + A +).

On lauseid, mis pole ei grammatiliselt korrektsed ega ka aktsepteeritavad

(G – A –).

Lause võib olla grammatiliselt korrektne, kuid sellegi poolest mitteaktsepteeritav

(G + A –).

Värvitud rohelised unenäod magavad raevukalt.

Lause võib olla ka grammatiliselt ebakorrektne, kuid siiski aktsepteeritav

(G – A +).

Tulen natuke peale kolme.

Vigu on võimalik analüüsida kas vea piirkonda või vea ulatust arvestades. (Ellis, Barkhuizen 2005: 59) Vea piirkond (*domain*) on fraas, sõna, osalause, lause, eelnev lause või diskursus, mida tuleb analüüsida, et viga identifitseerida (vt joonis 11). Vea ulatus (*extent*) näitab, missugust üksust tuleb vea parandamiseks muuta.

They **passed near a zoo and *stop*** in a forest.

Joonis 11. Näitelauses on veapiirkond tähistatud rasvase kirjaga, vea ulatus kaldkirjas (Ellis, Barkhuizen 2005: 59).

Vead võivad erineda ka raskusastme poolest. Corder (1981: 10) on veaks (*error*) pidanud ebaõiget keelekasutust, mida keelekasutaja ise parandada ei oska. Väiksemad vead, eksimused (*mistake*), on sellised, mida õppija suudab ka ise parandada, kui neile tähelepanu juhtida. Eksimused näitavad, mida õppija on küll ära õppinud, kuid ei suuda veel ise praktikas rakendada. Keskenduda tuleb vigadele, mida õppija ise parandada ei suuda.

Vead võivad olla selgesti nähtavad (*overt*) või varjatud (*covert*). Selgesti nähtavaid vigu on võimalik leida nende lähimat ümbrust analüüsides. Varjatud vigade leidmiseks tuleb analüüsida lõiku, tervikteksti või suulise kõne puhul kogu vestlust. (Ellis, Barkhuizen 2005: 56)

5.2. Veaanalüüsi etapid

Pit Corder (1974 viidatud Ellis, Barkhuizen 2005: 57 kaudu) on veaanalüüsi etappidena nimetanud

- materjali kogumist,
- vigade identifitseerimist,
- identifitseeritud vigade klassifitseerimist,
- vigade põhjuste seletamist;
- vigade raskusastme hindamist (vigadele hinnangu andmist) (*evaluation*), mida ei saa pidada vigade analüüsimise eraldiseisvaks etapiks, kuid mis siiski kuulub vigade analüüsimise juurde.

Traditsiooniline veaanalüüs (*error analysis*) hõlmab R. Ellise sõnul (Ellis, Barkhuizen 2005: 57) kolme etappi:

- vea tuvastamine (*identification*),
- vea kirjeldamine (*description)*,
- vea selgitamine (*explanation*).

Projektis *Free Text* kasutasid Dagneaux, Denness ja Granger õppijate vigade korpuspõhiseks analüüsimiseks metodoloogiat, mis sisaldas järgmisi etappe:

1. korpuse tekstis leiduvate vigade parandamine;
2. detailse veamärgendussüsteemi väljatöötamine;
3. veatähistuste ning paranduste sisestamine;
4. vigade statistika ning veatüüpide analüüs;
5. peamiste veatüüpide analüüs konkordantse uurides (Granger 2003: 466).

Esitatud veaanalüüsi etapid võib koondada ning loogilisse järgnevusse seada.

5.2.1. Tekstide kogumine, parandamine ja veaklassidesse liigitamine

Veaanalüüsi I etapp hõlmab õppijakeele näidete kogumist, tekstide parandamist ning paranduste põhjal veaklassifikatsiooni väljatöötamist.

R. Ellise sõnul teeb õppija vigu nii mõistmisel kui eneseväljendusel. Mõistmisvigade tuvastamine ja fikseerimine on vaieldav, kuid eneseväljendusel (nii suulisel kui kirjalikul) tehtud vigu on võimalik koguda ja uurida. (Ellis, Barkhuizen 2005: 51)

40 aastat tagasi analüüsis L. Villandi eesti õpilaste vigu sel moel, et kirjutas kirjanditest leitud vead perfokaardile ning moodustas sõnastusvigade kartoteegi, millest sai uurimistöö alus (Villandi 1966: 18).

Uuemat materjali kogumise viisi – õppija keelekorpuse loomist ning sellega seonduvaid aspekte, mis võivad vigade hulka mõjutada, tutvustati pikemalt esimeses peatükis.

Tekstides tuvastatud vead parandatakse ja tähistatakse. Objektiivsete tulemuste saamiseks on vaja ka sagedusanalüüsi, mille abil saab hinnata grammatiliste vormide ning keelendite üle- või alakasutust (vt lisa 4). Statistiline analüüs toob selgelt välja erinevatele žanritele või keelekasutajatele iseloomulikud tunnused ning võrdlusaluse olemasolul saab hinnata, kuidas kogumisel arvesse võetud tegurid tekstides kajastuvad (Granger 2003: 470). Sagedusanalüüsi tegemiseks ei ole vaja märgendatud korpust, küll aga peab korpus olema lemmatiseeritud.

Veaklassifikatsiooni ehk veataksonoomia loomisel tuleb lähtuda analüüsitavates tekstides sisalduvatest vigadest või analüüsida võrreldavate ja võrdlusaluseks olevate tekstide erinevusi ning selle põhjal vead kategoriseerida. Keelevigade abstraktsetesse kategooriatesse jagamine põhjustab suuri probleeme, kuna veaklassid ei ole selgepiirilised ning ühte viga on võimalik tõlgendada ja ka parandada mitmel viisil. "Ühe veaklassi piires ning veaklasside vahel moodustub vealiikide ja alamliikide vaheliste seoste võrgustik. Eeldatavalt võivad seosed olla korrelatiivsed, mitmesuunalised, üksteist ja/või teineteist välistavad, hajusad." (Eslon 2007: 108) Kuigi iga viga võib tõlgendada mitut moodi, peavad hilisema uuringu tegemiseks kõik

tekstikogus leiduvad vead olema tõlgendatud ühtemoodi. Juba üsna lühikese teksti katseline analüüsimine toob välja hulga probleeme, mis tuleb märgendajate vahel läbi arutada, et tulemus oleks võimalikult ühtlane (Meunier 2007: 27).

Korpusanalüüsis kasutatav vigade taksonoomia peab olema lihtne ja selge. Vigade liigitamisel on võimalik kasutada erinevaid süsteeme, näiteks määratleda vea raskusaste (üldine viga, spetsiifiline viga) või vea laad: (grammatiline, leksikaalne, leksiko-grammatiline, registriviga, süntaksiviga, stiiliviga vm (Granger 2007: 58–59).

Corder (1981: 38) ei pea piisavalt põhjalikuks klassifikatsiooni, mis jagab vead ortograafia, grammatika ning leksika tasandil lihtsalt väljajätte, lisamise, järjestamise ja valikuvigadeks, sest selline väliste tunnuste alusel tehtud klassifikatsioon ei võimalda sisulist analüüsi.

Dulay, Burt´i ja Krasheni veataksonoomia järgi võib vigu klassifitseerida kahte alternatiivsesse ja omavahel kombineeritavasse kategooriasse: lingvistilisse ja struktuuritaksonoomiasse. Need kaks taksonoomiat ei välista teineteist ning parimal juhul arvestab veakirjeldus mõlemaid (Granger 2003: 467, Ellis, Barkhuizen 2005: 51–73).

Lingvistiline taksonoomia põhineb kasutatava keele deskriptiivse grammatika üldistel kategooriatel nagu lausestruktuur, verbifraas, nimisõnafraas jne (üldisemalt jaotatud vigu võib vajadusel klassifitseerida ka detailsemalt).

Struktuuritaksonoomia loetleb neli peamist struktuuri muutmise viisi: väljajätmine; lisamine, väärmoodustus; vale järjestamine. Carl James (1998: 111) lisab viiendaks ka sulandid (ingl *blends*).

A. Krüger ja S. Hamilton (1997: 53) on veaparandusprogrammis *Repairing Errors in Computer Assisted Language Learning* (RECALL) klassifitseerinud keelevigu järgmiselt:

ortograafiavead (tähtede märkimatajätmine või lisamine);
morfoloogiavead (sõnavormivead, käänamine, pööramine);
süntaksivead (ühildumine, lausekonstruktsiooni vead, sõnade järjekord);
sõnavalikuvead (vale sõna kasutamine);
semantilised vead (idioomivead);
pragmaatilised vead (fraas või lause grammatiliselt õige, kuid pragmaatiliselt sobimatu). (Krüger, Hamilton 1997: 53)

Carl James (1998: 144–166) on produktsioonivead jaganud leksikaalseteks, grammatilisteks ja diskursuse vigadeks (James 1998: 144–166, R. Pajusalu, 2004: 218–235).

Leksikaalsed vead (leksikaalset tähendust omavate sõnade väärkasutus):

- formaalsed vead – vale moodustuse või sõnade segiajamise tõttu valitakse vale sõnavorm
- semantilised vead – valitakse vale sõna, mis on küll formaalselt õige, kuid ei oma seda tähendust, mida kõneleja on arvanud

Grammatilised vead:

- morfoloogilised vead;
- süntaktilised vead – fraasi, osalause ja lause tasandil tehtud vead;
- kohesioonivead.

Diskursuse vead:

- koherentsivead – teksti struktuuri seisukohalt tehtud vead, nt viitamisvead ja (eelkõige deiktilise tähendusega) sõnade väärkasutus;
- pragmaatilised vead – eelkõige sotsiokultuurilised vead nagu näiteks vale viisakusstrateegia valimine.

Tabel 7 esitab veaklassid erinevates veamärgendustega õppijakorpustes.

Tabel 7. Tähistatavad vead õppijakorpustes (Diaz, Fernandez 2006: 90).

	CLC	Free-Text	Louvain	NICT JLE (suulise kõne korpus)
fonoloogia	-	-	-	-
interpunktsioon	+	+	+	-
ortograafia	+	+	+	-
grammatika	+	+	+	+
leksika	+	+	+	+
pragmaatika	+ registrivalik	+ registrivalik	+ registrivalik	+ mitteverbaalne suhtlus
diskursus	asesõna viitamine	kohesioon	ebaloomulikud või võimatud väljendused	eneseparandused

R. Pool ja E. Vaimann on eesti keelt teise keelena kesk- ja kõrgtasemel õppijate vigu parandanud ning veaklassidesse jaganud. Arvesse on võtetud nii struktuuritaksonoomit (vale täht, liigne täht, poolik lause jne) kui ka lingvistilist taksonoomiat (sõnajärg, täis- ja osasihitis, vormimoodustus, käände valik jne). Eristus 25 vealiiki, millest sagedaminiesinevaid oli 14 (vt lisa 5) Vead on klassifitseeritud nende esinemissageduse järgi. (Pool 2007: 210)

Tallinna ülikooli eesti keele kui teise keele õppija korpuses on vead jagatud 18 veaklassi vahel (vt joonis 12). Veaklasside abstraktsioon tugineb "arusaamale keelesüsteemi paradigmaatilisest ja süntagmaatilisest ülesehitusest ning kujuneb nende kahe telje ristumispunktides. Paradigmaatilisteks tunnusteks on hierarhia grafeem–morfeem-sõna-sõnaühend-lause-tekst, mis süntagmaatilisel teljel on seostatud keelesüsteemi kolme aspektiga – semantika, grammatika ja pragmaatikaga." (Eslon 2007: 107)

Süntagmaatika / Paradigmaatika	Semantika	Grammatika	Pragmaatika
Tekst	1	2	3
Lause	4	5	6
Sõnaühend	7	8	9
Sõna	10	11	12
Morfeem	13	14	15
Grafeem	16	17	18

Joonis 12. Tallinna ülikooli vahekeelekorpuse keelevigade lingvistiline klassifikatsioon (Eslon, Metslang 2007: 101).

Emakeelt õppivate laste kirjandites esinevaid sõnastusvigu (vt lisa 7) on 1966. aastal klassifitseerinud Leo Villandi (1966). L. Villandi sõnul ei saa vigade liigitamise aluseks võtta ainult keele grammatilist struktuuri, kuna vead ei mahu "puhtteoreetiliste skeemide Prokrustese sängi: kuidas neid ka ei liigita, ikka tekivad kas puudu- või ülejäägid, ristumised või ebamäärasused" (Villandi 1966: 20). Ta jagas õppijate tekstides esinenud vead järgmiselt: sõnade tähendusvead; kordus ja paljusõnalisus, seosevead; argumenteerimta mõtteavaldus; vasturääkivus; liigitusvead; väljendusvahendite üksluisus; üldsõnalisus; isikupära puudumine; stiilipuhtuse vead; kujundilisuse vead, grammatilis-sõnastuslikud vead (Villandi 1966: 26–99).

K. Habicht on tänapäevast noortekeelt analüüsides välja toonud õpilastekstides sagedamini esinevad ortograafia-, vormi- ja stiilivead, kuid tema eesmärgiks ei ole olnud kõikide vigade klassifitseerimine, vaid sagedatele veatüüpidele osutamine (vt lisa 6). (Habicht 2008)

Kõikide tekstis esinevate vigade klassifitseerimiseks sobiva liigitusaluste süsteemi leidmine on komplitseeritud, kuid korpuse veamärgendussüsteemi loomiseks möödapääsmatu etapp.

5.2.2. Märgendite süsteemi väljatöötamine ja korpuse märgendamine

Veaanalüüsi II etapp

Enamasti on õppijakeelekorpused mahukad, sisaldades miljoneid sõnu. Nii suure materjali märgendamine on väga aeganõudev isegi juhul, kui on võimalik kasutada mingit vea tuvastamise programmi. Iga vea määramiseks tuleb arvestada mitmeid tegureid ning teha palju otsustusi (Ellis, Barkhuizen 2005: 241). Märgendatus on siiski oluline, kuna annab märgendamata tekstiga võrreldes oluliselt rohkem võimalusi. Märgendusest on kasu ka juhul, kui see on lisatud vaid ühele osale korpuse tekstidest.

Vigade märgendamise juures on kõige keerulisem klassifitseerimise aluse leidmine. Klassifitseerimisalus peab hõlmama võimalikult paljusid esinevaid vealiike ning võimaldama ühele sõnale (fraasile, lausele, lauseosale) lisada vajadusel ka mitu erinevat märgendust. Ükskõik, milline süsteem kokku lepitakse, kuid kõik töötajad peavad seda ühtemoodi interpreteerima (Bowker, Pearson 340), mistõttu on väga oluline, et kõik veatüübid ja nende märgendamise alused defineeritaks selgelt ja dokumenteeritaks korralikult.

S. Grangeri (2003: 466) hinnangul peab efektiivne vigade märgendamise süsteem olema

1. Informatiivne (ingl *infomative*) ja hallatav (ingl *manageable*) – piisavalt detailne, et anda õppija vigade kohta vajalikku informatsiooni, kuid mitte nii detailne, et märgendaja ei suuda seda hallata.

2. Korduvkasutatav (ingl *reusable*)– kategooriad peavad olema piisavalt üldised, et neid oleks võimalik erinevate keelevariantide analüüsimisel rakendada

3. Kohandatav (ingl *flexible*)– märgendamise ajal peab olema võimalik märgendeid lisada või kustutada, pärast annotatsiooni mitmesuguseid otsinguid teha.

4. Järjekindel ja terviklik (ingl *consistent*) – et kindlustada maksimaalset järjekindlust annoteerijate vahel, peab koostama vigade märgendamise eeskirjad. (Granger 2003: 467)

Soovitatud on, et sama teksti märgendaks vähemalt kaks inimest, kes saavad omavahel märgendusi võrrelda ning tekkinud erimeelsused üle vaadata (Kaalep 2008 eravestlus). Kuigivõrd jääb vigade tõlgendamine siiski subjektiivseks. Mida detailsemat süsteemi kasutatakse, seda rohkem erinevast tõlgendusest tulenevaid eksimusi teevad märgendajad. (Meunier 2007: 27)

Nijmegeni ülikoolis Hollandis on välja töötatud vigade märgendamise süsteem (ingl *Error editor*), mis aitab veamärgendeid tekstis õigesse kohta lisada. Eristatakse seitset veakategooriat: formaaalne, grammatiline, leksiko-grammatiline, leksikaalne, registriviga, stiiliviga ning viga, mille puhul on sõna puudu / sõna üle / sõnad vales järjekorras. (Pravec 2002: 98)

Louvainis välja töötatud hierarhilises vigade märgendamise süsteemis määratakse iga vea puhul selle üldkategooria (Bowker, Pearson 2002: 340) ning lisaks üldkategooriale määratletakse kõik vead ka spetsiifilisemalt. Iga vealiik on defineeritud ning selle kohta on esitatud näide. Üldkategooriateks on grammatiline (G), leksikaalne (L), leksikogrammatiline (X), formaalne (F), registri- (R), süntaksi- (W) ja stiiliviga (S). (Ellis, Barkhuizen 2005: 241)

James (1998) soovitab liita veaklassid üheks kahe- või isegi kolmedimensiooniliseks taksonoomiaks, et võimaldada erineva spetsiifilisuse astmega otsinguid. Prantsuskeelse korpuse FRIDA annotatsioon on kolmedimensiooniline. Kokkuleppeliselt ei eristata selles vigu ja eksimusi ega ka inter- ja intralinguaalseid vigu. Märgenduste sisestamiseks on kirjutatud vastav programm. (Granger 2003: 470)

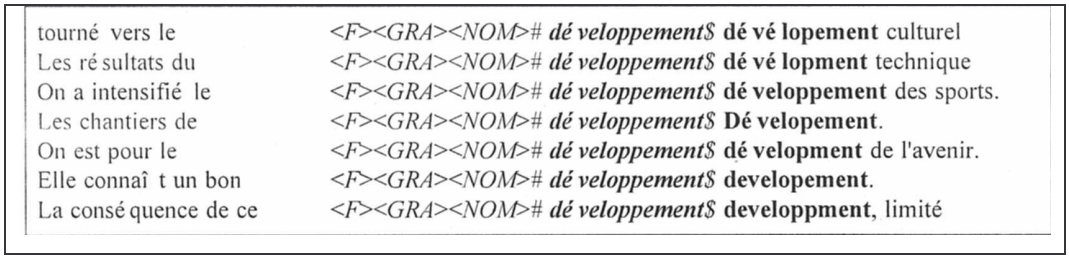
Kolm märgendatud valdkonda on:

I – vea domeen (mõjupiirkond) ja kategooria. Vea domeen on kõige üldisem tasand, mis näitab, kas viga on formaalne (näiteks ortograafiline), grammatiline,

leksikaalne vm. Domeene on üheksa. Kõik domeenid on jagatud veakategooriateks. (Granger 2003: 470)

II – sõnaliigid. Vigased kohad on morfoloogiliselt märgendatud, mistõttu on võimalik eristada vigu sõnaliigiti ja tuua välja tüüpilised veaohtlikud kohad. Õppijakorpuse morfoloogilisel märgendamisel ei saa kasutada programmi, kuna programm ei pruugi vigadega teksti õigesti tõlgendada. (Granger 2003: 470)

III – parandused. Vigase vormi järele on sisestatud korrektne vorm. Paranduste lisamine on oluline kahel põhjusel 1) võimaldab vea annotatsiooni paremini interpreteerida; 2) võimaldab korrektsete vormide põhjal leida sama sõna erinevaid vigaseid versioone (vt joonis 13). (Granger 2003: 470)



Joonis 13. Õppijakorpuses esinevate ortograafiavigade esitlemine korrektsete vormide põhjal (Granger 2003: 470).

Vigade märgendamiseks peab märgendaja oma intuitsiooni süstematiseerima. Mõne märgendussüsteemi puhul tuleb annoteerijal teha valikuid võimalike veakategooriate ning neis sisalduvate alatüüpide vahel, teisel juhul peab annoteerija mitmest erineva tähendusega sümbolist ise veatähise moodustama või ühele veale mitme erineva taseme annotatsiooni lisama. (Diaz, Fernandez 2006: 89)

Olenevalt tekstidele lisatud metaandmetest on võimalik analüüsida vigu kas üldisemalt või keskenduda iga konkreetse õppija vigadele eraldi. Kui korpuses esinevatele vigadele on korrektsed keelendid juurde kirjutatud, sobib korpus hästi ka keeleõppijale. (Bowker, Pearson 2002: 340) Prantsuse keele õppijakorpuses FRIDA on vigade analüüs automatiseeritud. Õppija teksti vigaste keelendite juurde lisatakse automaatselt viited ka virtuaalse grammatikaõpiku vastavale peatükile.

Milton ja Chodwhury on inglise keelt õppivate hiinlaste korpuses pakkunud parandustena välja mitut tõlgendust (Ellis, Barkhuizen 2005: 241). Ka Saksamaal Humboldti ülikoolis loodud *Fehler-Annotierte Lingvistische Korpus* (FALCO) esitab mitme parandamisvõimaluse korral kõik parandamise võimalused (vt joonis 14) (Lemnizer 2006: 88).

Token	**Die**	**Erkärung**	**für**	**diese**	**Phänomen**
ZIEL 1				dieses	
Fehler				Genus	
ZIEL 2					Phänomene
Fehler					Numerus

Joonis 14. Õppijakorpuse FALCO veaparandussüsteem mitme parandusvõimalusega vea puhul (Lemnizer 2006: 88).

5.2.3. Korpuse analüüs

Veaanalüüsi III etapp

- märgendatud korpuse vigade statistika ning veatüüpide analüüs;
- vigade kirjeldamine ja selgitamine;
- õppija vigade kaalukuse hindamine,

Märgendatud korpuse vigade statistika ning veatüüpide analüüs

Veamärgenditega korpuse statistiline analüüs esitab vead tüübiti, mis on oluline veatüüpide kohta käivate üldistavate järelduste tegemiseks. Ellis (2005: 55–56) väidab, et kuigi veamärgendite sisestamine on aeganõudev, sisaldab märgenditega tekst väärtuslikku informatsiooni õppija keele kohta. Ingliskeelsete õppijatekstide veaanalüüs on näidanud, et nii emakeele- kui ka teise keele õppijate vead on regulaarsed ja üldistatavad. Kuna õppija hakkab kasutama järjest keerulisemaid konstruktsioone, võib tekstis esinevate vigade koguhulk vanematel koolilastel olla isegi suurem kui nooremas klassis õppijatel. (Ellis, Barkhuizen 2005: 55–56) Vigade iseloom on kogu õpiaja jooksul selge kuluga, st vead arenevad, muutuvad. Üksikasjalik ja detailne õppijakeele vigade kirjeldus aitab õppijal saavutada keelekasutuse korrektsust. (Granger 2003: 466)

Vigade kirjeldamine ja selgitamine

Õppijate vigade juures väärib tähelepanu nende süstemaatilisus (Ellis, Barkhuizen 2005: 55), mistõttu on vigu võimalik ka üldistavalt kirjeldada. Vigade selgitamisel analüüsitakse põhjusi, miks viga on tehtud.

Vigade põhjusi saab jagada näiteks interlinguaalseteks ja intralinguaalseteks. Interlinguaalsed vead on tekkinud teise keele mõjutusel. Võõrkeeleõppija puhul on õpitava keele mõjutajaks õppija emakeel ning teised varemõpitud keeled. Emakeelset teksti kirjutava õppija keelekasutust võib mõjutada kohalik murdekeel, mõni õpitav (või populaarne) keel. (Corder 1981: 15–16)

Intralinguaalsed vead peegeldavad õppimisstrateegiat ning on kõikide keelte puhul ühesugused. C. James (1998: 184–187) eristab seitset tüüpi intralinguaalsete vigade põhjusi:

- väär analoogia (ingl *false analogy*);
- puudulik analüüs (ingl *misanalysis*);
- reegli mittetäielik rakendamine või valesti tõlgendamine (ingl *incomplete rule application*);
- liiasus (ingl *exploiting redundancy*);
- ei nähta sõnade koosesinemise võimalikke piiranguid (ingl *overlooking cooccurence resrictions*);
- õige parandamine valeks (ingl *hypercorrection, monitor overuse*);
- liigne üldistamine, süsteemi lihtsustamine (ingl *overgeneralization, system-simplification*).

Ellis hindab vigade tekkimise mõjureid õppijast, keelest ning väljundist lähtuvalt. Spetsiifiliste tulemuste saamiseks kogutakse ainult selliseid tekste, mis vastavad väga täpselt kogumisnõuetele. Kui eesmärgiks on üldisem keelekasutuse uurimine, võib tekste koguda suurema varieeruvusega. (vt pt 2). (Ellis, Barkhuizen 2005: 57)

Õppija vigade kaalukuse määramine (ingl *error evalution*)

Vea hindamine on veaanalüüsi lisaetapp, millel on eelkõige pedagoogilised eesmärgid. Mida suurem on otsustajate hulk, seda usaldusväärsema tulemuse vigade

hindamine annab. Emakeeleõppijate vigu tuleb hinnata mitte-emakeelsete õppijate omadest eraldi. (Ellis, Barkhuizen 2005: 51)

Hinnatavad vead esitatakse enamasti täislauses või tekstis ning palutakse määrata vea kaal skaalal, mille ühes otsas on raske viga ja teises keelend, milles ei ole viga. (Ellis, Barkhuizen 2005: 51) Vigade raskusastme määramisel hinnatakse õppija vigade suhtelist raskusastet ning hindaja lähtub oma vaatepunktist. Rääkimata aktsepteerimisel põhinevast stiilivigade hindamisest, võib probleemne olla ka näiteks kirjakeele normist lähtuvate vigade kaalu määramine. Eesti keeleõppijate ortograafia- ja morfoloogiavigade hindamisest on kirjutanud U. Kuresoo (1990: 462–470). Eksperimendi korras analüüsiti 130 gümnaasiumiõpilase testilaadseid ortograafia- ja morfoloogiaalaseid teste. Seejärel paluti õppijate vigu hinnata emakeeleõpetajatel ning TÜ õppejõududel ja keeleteadlastel. Õpetajate hinnangul oli viga raske 101 juhul ning õppejõudude hinnangul vaid 6 juhul. (Kuresoo 1990: 167–168)

R. Pajusalu on analüüsinud 20 eesti keelt võõrkeelena õppinud vene emakeelega TÜ lõpetaja lõpukirjanditest leitud vea hindamist ning leidnud, et sageli ei ole hindajad hinnangut andnud mitte veale, vaid kogu lausele (Pajusalu 2004: 231).

Veaanalüüs annab õppija keelest mõneti moonutatud pildi, kuna analüüs on ühekülgne: esile tõstetakse vaid vead ning tähelepanuta jäävad korrektsed keelendid, huvitavad lausekonstruktsioonid ning mitmekesine sõnavara (Ellis, Barkhuizen 2005: 70). Sobiva märgendamissüsteemi olemasolu võimaldab analüüsida ka teisi huvipakkuvaid keelenähtusi.

Keelevigu sisaldav korpus on vajalik grammatikakorrektori arendamiseks (Kaili Müürisep 2008, eravestlus).

6. Eesti kooliõpilaste kirjalik tekstikorpus

Magistritöö praktiline osa on jagatav kahte alapeatükki: emakeeleõppija korpuse kirjeldus ja statistiline analüüs ning emakeeleõppija korpuse veaklassifikatsiooni aluste ja veamärgendusjuhendi koostamine.

6.1. Emakeeleõppijate tekstikorpus

6.1.1. Korpuse tekstid

Eesti koolilapse keelekasutuse uurimiseks olen koostanud tekstide kogu, mida võib nimetada emakeeleõppija korpuseks. Emakeeleõppija korpuse eesmärk on koguda ainest õppijate keelekasutuse analüüsimiseks. Korpus pakub võimaluse uurida õppijate tekste korpusanalüüsi vahenditega.

2008. aasta mais sisaldab emakeeleõppija korpus esimesi tekste. Nende kogumisel on arvesse võetud autorite emakeelt, vanuselist ja piirkondlikku eripära, teksti tüüpi ja kirjutamise situatsiooni. Teksti karakteristikute ja kirjutamise situatsiooni põhjal võib kogutud materjali liigitada kolme gruppi: eksamikirjandid, koolikirjandid ja võistluskirjandid.

Eksamikirjandid

Alates 1997. aastast on igal kevadel Riikliku Eksami- ja Kvalifikatsioonikeskuse (REKK) arhiividesse laekunud u 10 000 eksamikirjandit nii gümnaasiumi, kutsekooli kui ka õhtukooli lõpetajatelt. Kirjandid on kirjutatud kontrollitud situatsioonis, abivahendina on lubatud kasutada õigekeelsussõnaraamatut. Eksamikirjandi peavad kirjutama kõik põhikooli või gümnaasiumit lõpetavad õpilased, mistõttu kirjandite tase erineb väga suurel määral (veatutest kirjanditest 60–70 veaga kirjanditeni) (Hennoste 2008, eravestlus). REKK ei piira kirjandite kasutamist uurimise eesmärgil, kuid elektroonilise eksamikirjandite kogu loomine on raskendatud – kirjandite hulk on küll tohutult suur, kuid need kõik on käsikirjalised. Põhikooli lõpus kirjutatavaid kirjandeid hinnatakse ja säilitatakse koolides.

Emakeeleõppija korpusesse kogutud eksamikirjandite hulk on väike. Kirjandid on pärit M. Rõigase koostatud kirjandiõpetuse kogumikest ning on sinna tükitud originaaltekstide järgi. (Madissoo 1998 ja Madissoo 2002). Leo Villandi sõnul (1966: 17) rakendavad õpilased eksamil kogu oma vaimujõudu, mistõttu lopsakaid vigu leidub neis töödes vähem. Arvestada tuleb ka eksamisituatsioonile omast pingelist õhkkonda. (Villandi 1966: 17) Teiste emakeeleõppija korpuses sisalduvate tekstidega võrreldes on eksamikirjandid kirjutatud ajaliselt mõned aastad varem – 1997. ja 2001. aasta kevadel.

Koolikirjandid

Koolikirjanditeks nimetan kirjandeid, mida õppija on kirjutanud koolis õpetaja antud teemal, kuid mis ei ole kirjutatud eksamisituatsioonis (ka *jooksva töö kirjandid* Villandi 1966: 17). Õpilane võib olla kasutanud abivahendeid. Koolikirjandeid esitavad õppijad ka elektrooniliselt ning nende kogumine on seetõttu eksamikirjanditega võrreldes hõlpsam. Koolikirjandeid kirjutavad kõik õppijad.

Emakeeleõppija korpuse koolikirjandite kogu sisaldab 2008. aasta kevadel 115 eesti emakeelega gümnaasiumiõpilase kirjandit, mis on kirjutatud 2007.–2008. aastal. Kirjandid on kogutud kolme maakonna viiest koolist. Enim tekste on pärit Viljandi maakonnast. Tekstid on kogutud emakeeleõpetajate abiga.

Võistluskirjandid

Kolmas liik kirjandeid esindab kirjutajate paremikku, kuna tööd saadetakse teistega konkureerima. Võistluskirjandi kirjutamiseks võib õppija nõu küsida õpetajalt või teistelt võimalikelt nõuandjatelt. Võistluskirjandite kogumine on teiste kirjanditega võrreldes kõige lihtsam, sest need esitatakse elektroonilisel kujul (perioodil 2004–2007 oli kaks võistluskirjandit esitatud käsikirjalisena).

Emakeeleõppija korpuse võistluskirjandid on pärit SA Innove korralikult dokumenteeritud arhiivist. Kirjandite koguhulgast on eraldatud vene emakeelega õppijate kirjandid. Esialgne arvamus, et võistluskirjanditena laekub vaid keeleliselt väga korrektseid kirjandeid, ei pidanud paika.

Põhikooliõpilaste kirjandid

Põhikooliõpilaste kirjandite kogu ei ole homogeenne, sisaldades nii kooli- kui ka eksamikirjandeid aastatest 2005–2008. Kuna põhikooliõpilastelt kogutud tekstide koguhulk oli väike, ei peetud otstarbekaks eksamikirjandeid koolikirjanditest eraldi analüüsida. Metaandmete põhjal on võimalik eksamikirjandeid koolikirjanditest eristada ning korpuse suurenedes tulebki seda teha. Põhikooliõpilaste kirjandid on kogutud neljast Tartu linna ja maakonna koolist põhikooli vanemast osast ehk 7.–9. klassi õpilastelt. Kirjandid on kogutud emakeeleõpetajate abiga.

Emakeeleõppija korpus sisaldab 2008. aasta kevadel

61 793 sõna gümnaasiumiõpilaste koolikirjandite tekste;

18 763 sõna gümnaasiumiõpilaste eksamikirjandite tekste;

243 719 sõna gümnaasiumiõpilaste võistluskirjandite tekste;

12 114 sõna põhikooliõpilaste tekste.

6.1.2. Metaandmed

Iga korpusesse lisatud teksti ees on teksti autori ja teksti tüübi kohta käivad metaandmed. Metaandmed on esitatud nurksulgudes.

1. Teksti tüüp. Tähistused on loodud ainult nende tekstitüüpide tarvis, mis praegu korpuses olemas on. Kasutatud on järgmisi tähiseid: eksamikirjand <EKS>, koolikirjand <KOK>, võistluskirjand <VST>.

loovtöö liik	**tähis**
eksamikirjand	<EKS>
koolikirjand	<KOK>
võistluskirjand	<VST>

2. Õpilase vanuseaste

Õpilase vanuseaste on määratletud kooliastmeti. Kogutud tekstide autorid õpivad kolmandal kooliastmel (<PV>) või gümnaasiumis (<GÜ>).

kooliaste	õppija vanus	tähis
põhikooli vanem vanuseaste III kooliaste (7.–9. klass)	14–17 aastat	<PV>
gümnaasium (10.–12. klass)	17–19 aastat	<GÜ>

3. Teksti kirjutamise aeg

Teksti kirjutamise aeg tähistatakse aastaarvuga. Praegu kogus sisalduvad põhikooliõpilaste tekstid on aastatest 2005–2008, koolikirjandid aastast 2007–2008, võistluskirjandid aastatest 2004–2007 ja eksamikirjandid aastatest 1997 ja 2001.

4. Kirjutaja asukoht

Märgendid on loodud ainult nende kohtade tähistamiseks, kust tekste on kogutud. Näiteks koolikirjandi autorite asukohta tähistavad kolmest sümbolist loodud tähised: Otepää linnast ja maakonnast pärit tekstid <OTP>, Tartu linnast ja maakonnast pärit tekstid <TRT> ning Viljandi linnast ja maakonnast pärit tekstid <VIL>.

kirjutaja asukoht	koha tähis korpuses
Otepää	<OTP>
Tartu	<TRT>
Viljandi	<VIL>

5. Identifitseerimisnumber

Igal emakeeleõppija korpusesse lisatud tekstil on identifitseerimisnumber, mille järgi on võimalik seda teistest samasse gruppi kuuluvatest tekstidest eristada.

6. Õpilase emakeel

Õpilase emakeelt ei ole tähistatud, kuna kogu sisaldab praegu ainult eesti emakeelega õppijate tekste. Vajadusel on võimalik vastav tähis automaatselt lisada.

Esialgu oli kavas tekstide juurde märkida ka kirjutaja sugu, kuid kirjanditele ei ole seda infot järjekindlalt lisatud, kuna kõikide kirjutajate kohta ei olnud nende sugu teada.

Näide teksti metaandmete kohta:

<KOK/GÜ/2007/VIL/1>

(Viljandi maakonnast pärit gümnaasiumiõpilase 2007. aastal kirjutatud koolikirjand).

6.2. Emakeeleõppija korpuse statistiline analüüs

Emakeeleõppija korpuse üldandmeid on analüüsitud statistiliselt, hinnates õppijateksti sõnade ja lausete keskmist hulka ning keskmist lausepikkust.

6.2.1. Üldandmed

Analüüs kajastab 578 kirjandi andmeid, mis sisaldasid kokku 26 537 lauset ehk 336 389 sõna (vt tabel 8). Gümnaasiumiõpilaste koolikirjandid sisaldasid keskmiselt 40 lauset (537 sõna), eksamikirjandid 43 lauset (569 sõna), kuid võistluskirjandid olid pikemad – keskmiselt 50 lauset (630 sõna). Põhikooliõpilaste tekstide pikkuseks oli keskmiselt 26 lauset (282 sõna).

6.2.2. Lausepikkus

Õpikute keerukuse hindamisel on üheks teksti keerulisuse faktoriks peetud lausepikkust ning lähtutud seisukohast, et põhikooliõpilase jaoks mõistetav lause ei tohi 7. klassis ületada 13 sõna, 8. klassis 14 sõna ning 9. klassis 15 sõna (Asser *et al.* 2004: 82). 1990. aastate ajakirjandustekstide lausepikkuseks on 13,06 sõna ning ilukirjandustekstides 10, 38 sõna (Kerge 2002: 44).

Emakeelt õppivate laste tekste analüüsides selgus, et gümnaasiumiõpilase koolikirjandi ja eksamikirjandi keskmine lausepikkus – 13,2/13,1 sõna – on 1990. aastatel ajakirjanduses kasutatud lausetest (13,06 sõna) pikem, kuid vastab ümardatult

20. sajandil eestikeelses ilu- ja ajakirjanduses kasutatavale kesmisele lausepikkusele, milleks on K. Kerge andmetel 13,1 sõna (Kerge 2002: 44). Õpilaste võistluskirjandite laused on veidi lühemad, sisaldades keskmiselt 12,5 sõna.

Anne Uuseni (2006: 148) uurimuse põhjal on 3. klassis õpilase lause optimaalseks pikkuseks 9–12 sõna ja 6. klassi õpilastel 10–13 sõna. Emakeeleõppija korpuse põhjal võib väita, et põhikooli vanemas astmes õppivate laste (7.–9. klass) tekstide keskmiseks lausepikkuseks on 10,9 sõna.

Tabel 8. Emakeeleõppija korpuse statistilised andmed.

tekstitüüp	sõnade arv	kirjandite arv	lausete arv	lause keskmine pikkus sõnades	lausete hulk ühes tekstis keskmiselt	teksti keskmine pikkus sõnades
gümnaasiumiõpilaste koolikirjandid	61793	115	4653	13,2	40	537
gümnaasiumiõpilaste eksamikirjandid	18763	33	1432	13,1	43	569
gümnaasiumiõpilaste võistluskirjandid	243719	387	19348	12,5	50	630
põhikooliõpilaste kirjandid	12114	43	1104	10,9	26	282
kokku	**sõnu 336389**	**kirjandeid 578**	**lauseid 26537**	kõikide õpilastekstide keskmine lause-pikkus **12,4** gümnaasiumi-õpilaste keskmine lausepikkus **12,9**	kõikides õpilas-tekstides lauseid keskmiselt **39,75** gümnaasiumi-õpilaste tekstides lauseid keskmiselt **44,3/**	kõikides õpilas-tekstides keskmine pikkus **504** sõna gümnaasiumi-õpilaste tekstide keskmine pikkus **578,6** sõna

6.2.3. Sagedusanalüüs

Emakeeleõppija korpuse sagedusanalüüsi peatükis esitatakse sagedamate kollokatsioonide ja sõnavormide loetelu ning esitletakse õppijatekstides sagedamini kasutatavaid sõnaliike: verbe, nimisõnu ja asesõnu. Võrdlusalusena kasutatud "Eesti kirjakeele sagedussõnastikus" on sagedusloetelu koostatud oluliselt suurema hulga (1 007 000 *versus* 336389) tekstisõnade põhjal, kuid võrdlus on siiski võimalik, sest analüüsitakse sõnade esinemist üksteise suhtes, mitte ei võrrelda absoluutarve.

6.2.3.1. Õppijatekstides leiduvad sagedamad kollokatsioonid

Eesti keelt emakeelena õppijate eksperimentaalkorpuse põhjal tehtud automaatne kollokatsioonide analüüs esitas sagedamatena peamiselt grammatilisi koosesinemisi (vt tabel 9). Põhikooliõpilaste tekstides esinenud mitte-grammatilised kollokatsioonid tulenevad õppijatekstide pealkirjadest "Mis huvitab tänapäeva noort" ja "Võit iseenda üle". Kollokatsioonide sageduse leidmisel on tekstide pealkirjad sagedusanalüüsist küll välja jäetud, kuid kuid põhikoolilapsed on oma tekstides kasutanud pealkirjas esitatud sõnu ja sõnaühendeid.

Tabel 9. Emakeeleõppija tekstides esinevad sagedamad kollokatsioonid.

kollokatsioonid põhikooliõpilaste tekstides	kollokatsioonid gümnaasiumi-õpilaste koolikirjandites	kollokatsioonid gümnaasiumi-õpilaste eksami-kirjandites	kollokatsioonid gümnaasiumi-õpilaste võistlus-kirjandites	kõik kollokatsioonid kokku
41 ja on	263 on ja	93 on ja	925 on ja	1220 on ja
32 on ka	221 et on	79 ja on	830 ja on	1087 ja on
32 on ja	216 ja on	67 et on	787 et on	1033 et on
32 ja et	171 on ka	46 on et	695 ei ole	874 ei ole
30 huvitab tänapäeva	157 ei ole	44 on ka	576 see on	735 see on
28 on on	152 on et	40 see on	549 on et	719 on et
27 mis tänapäeva	146 on on	40 oma ja	508 on on	688 on ka
26 ta ja	139 see on	40 ei ole	506 et ei	682 on on
25 on see	128 ja ei	39 mis on	501 ma ei	624 et ei
25 iseenda üle	126 mis on	33 ja ja	485 on ka	612 on see
25 et on	124 kui on	32 ta on	471 on see	594 mis on
24 oli ja	118 on väga	32 on on	447 mis on	574 kui on
23 tänapäeva noort	116 on see	32 ja ei	441 oma ja	568 ja ei
23 ma ei	113 ja ka	30 on see	440 kui on	550 ma ei
23 ja kui	113 ei saa	30 kui on	419 ja ei	540 oma ja

6.2.3.2. Õppijatekstide sagedamad sõnavormid

Eesti emakeeleõppija korpuse sõnavormide sagedusanalüüsi põhjal (vt tabel 10) on eesti õpilaste tekstides ootuspäraselt sagedaimad grammatilised sõnad. Silma jääb, et gümnaasiumiõpilaste kooli- ning eksamikirjandi tekstides ei ole sagedamate sõnavormide seas asesõna *ma* ning pronoomen *nad* esineb sagedana ainult põhikooliõpilaste tekstides.

Tegusõnadest on kõikides uurimisalustes allkorpustes esikohal olevikuline tegusõnavorm *on*. Minevikuline *oli* esineb põhikooliõpilaste tekstides ning gümnaasiumiõpilaste eksamikirjandites. Eksamikirjandite sõnavormide seas eristub ka modaalverbi vorm *võib*. Rohkem verbe 20 sagedaima sõnavormi seas ei ole. Nimisõnadest esineb gümnaasiumiõpilaste eksamikirjandites ning võistlustekstides 20 sagedaima sõnavormi seas *elu*.

Tabel 10. Emakeeleõppija korpuse sagedaimad sõnavormid

Tähised: **x** **verbivorm**
X **pronoomenivorm**
X **substantiivivorm**

põhikooli-õpilaste tekstid		**gümnaasiumi-õpilaste koolikirjandid**		**gümnaasiumi-õpilaste eksamikirjandid**		**gümnaasiumi-õpilaste võistluskirjandid**		**kõik sõnavormid kokku**	
445	ja	2318	on	690	on	8208	on	10848	on
322	on	1744	ja	592	ja	6970	ja	9159	ja
169	et	1007	et	277	et	4332	ei	5472	ei
169	ei	971	ei	265	ei	4170	et	5346	et
157	ma	770	ka	203	kui	2736	kui	3598	kui
143	kui	719	kui	203	ka	2449	oma	3307	ka
133	ta	678	ning	174	oma	2421	ka	3039	oma
126	oli	554	see	157	see	2235	ma	2934	ning
116	ka	497	oma	145	mis	2234	see	2898	see
110	see	396	mis	140	ta	2172	ning	2547	ma
93	oma	346	ta	138	ning	1632	mis	2113	mis
88	nad	325	kuid	112	nii	1374	aga	1757	aga
85	mis	324	nii	106	aga	1276	või	1631	nii
84	ning	314	seda	93	**elu**	1243	nii	1618	kuid
76	aga	307	aga	92	seda	1225	siis	1590	või
72	siis	277	siis	88	kuid	1222	kuid	1574	siis
71	kuid	275	kes	87	siis	1195	seda	1554	seda
64	nii	268	või	82	oli	1117	**elu**	1504	ta
56	väga	254	väga	79	kes	1103	kes	1433	kes
55	kes	252	palju	69	võib	1025	ta	1303	**elu**

Sõnavormide sagedusanalüüsi võrdlusalusena on võimalik kasutada eesti vahekeelekorpuse (Eslon 2007: 109–110) ja "Eesti kirjakeele sagedussõnastiku" (Kaalep, Muischnek 2002) sagedamaid sõnavorme. Õpilaste tekstidest lisasin võrdlusesse kirjandite koondi ning põhikooliõpilaste sagedamad sõnavormid (vt tabel 11). Jääb silma, et põhikooliõpilaste tekstide sõnavormid sarnanevad kirjakeele sagedamatega hoopis enam kui kirjandite koondi sõnavormid.

Esimesed neli sagedamat sõnavormi *ja*, *ei, et* ja *on* on väikese sageduserinevusega kõigis neljas allkorpuses. Esimene pronoomen on vahekeeleõppijatel ja põhikooliõpilastel *ma*, kirjandite koondis *oma* ning kirjakeeles *ta*. Kõige rohkem pronoomenivorme on vahekeelekorpuse tekstides, s.h isikulised asesõnad *ma*, *ta*, *nad* ja *me*. Emakeeleõppijate koondtekstide isikulised asesõnad on *ma* ja *ta*. Põhikooliõpilastel ka *nad*. Kirjakeele sagedamate seas eristuvad *ma*, *ta, tema* ja *nad*.

Minevikulist verbivormi *oli*, mis vahekeelekorpuse ning kirjakeele sagedussõnastiku põhjal on sageduselt 6. sõnavorm ning põhikooliõpilastel 8., ei ole õppijakorpuse 30 sagedama sõna seas üldse. Seevastu on õppijakorpuses sagedad *olen* ja *da*-tegevusnimi *olla.* H. Meier on esseetekste uurides leidnud, et kõige olevikulisem on hobide ja harrstuste valdkond ning esseedes kasutatakse minevikulist vaatepunkti vähem kui ilukirjanduses. (Meier 2003: 121)

Sidenditest jääb silma, et rinnastav sidesõna *ning*, mis eesti keeleõppijatel on sageduselt 8., põhikoolilastel 14. ja kirjakeele sagedussõnastikus 13., jääb vahekeelekorpuses alles kolmanda kümnendi lõppu.

Kirjakeelekorpuse analüüsis sagedamate sõnavormide seas ühtegi nimisõna ei eristu. Vahekeelekorpuses on sage pärisnimi *Eesti*, emakeeleõppijate korpuses nimisõna *elu* ja põhikoolis õppijatel *tänapäeva* (tõenäoliselt mõjutatud pealkirjast "Mis huvitab tänapäeva noort").

Tabel 11. Emakeeleõppija korpuse sagedamad sõnavormid vahekeelekorpuse (Eslon 2007: 109–110) ning kirjakeele sagedamatega (Kaalep, Muischnek 2000) võrreldes.

Tähised: x verbivorm
X pronoomenivorm
X substantiivivorm

vahekeeles	**emakeeleõppijate koond**	**kirjakeeles**	**põhikooliõpilaste kirjandid**
13939 ja	10848 on	27214 ja	445 ja
13295 on	9159 ja	19182 on	322 on
5553 et	5472 ei	13810 ei	169 et
5091 ei	5346 et	12314 et	169 ei
5087 ma	3598 kui	10170 ta	157 ma
4073 oli	3307 ka	8861 oli	143 kui
3882 Eesti	3039 oma	8599 kui	133 ta
3835 see	2934 ning	6191 ka	126 oli
3556 kui	2898 see	6114 see	116 ka
2991 ka	2547 ma	5329 oma	110 see
2478 oma	2113 mis	5274 aga	93 oma
2340 aga	1757 aga	4454 ma	88 nad
2337 ta	1631 nii	4409 ning	85 mis
2230 aga	1618 kuid	4391 mis	84 ning
1933 mis	1590 või	4238 siis	76 aga
1856 minu	1574 siis	3523 nii	72 siis
1741 palju	1554 seda	3310 või	71 kuid
1376 kõik	1504 ta	3177 seda	64 nii
1290 nad	1433 kes	2961 kes	56 väga
1260 sest	1303 elu	2961 nagu	55 kes
1246 või	1167 mida	2720 tema	47 sest
1243 nii	1142 minu	2550 veel	46 või
1233 siis	1096 palju	2534 pole	45 seda
1197 me	1076 sest	2341 kuid	44 palju
1189 olid	1066 kõik	2337 selle	42 olid
1182 seda	1029 olen	2287 kas	39 üle
1147 mida	1013 selle	2234 juba	37 kõik
1073 mulle	1011 väga	2190 nad	36 tänapäeva
1058 selle	988 pole	2008 välja	34 minu
1041 ning	943 olla	1978 midagi	33 peale

6.2.3.3. Kahekümne sagedasema lemma analüüs

Eesti emakeeleõppija korpuse sõnasagedusanalüüs on koostatud ka lemmatiseeritud tekstidele (vt tabel 12). Sagedamad lemmad on leitud samal moel nagu eesti kirjakeele sagedussõnastikku kuuluvate tekstide lemmad (Kaalep, Muischnek 2002: 11–13). Sagedusloetelu koostamisel on kirjandite pealkirjad statistikast välja jäetud. Kirjandite sagedusloetelu võrdlusaluseks on eesti kirjakeele sagedamad sõnad "Eesti kirjakeele sagedussõnastiku" andmetel (Kaalep, Muischnek 2002).

6.2.3.4. Asesõnad

Kõikides analüüsitavates kirjandikogudes on 20 sagedaima lemma seas enim asesõnu – koolilaste tekstide sagedaimate lemmades seas 7 (*mina, see, tema, oma, ise, kes, mis*) ja kirjakeeles 8 (*tema, see, mina, oma, sina, ise, kes, mis.*)

6.2.3.5. Sidesõnad

Sageduselt teisel kohal on sidesõnad: põhikoolikirjandites, kooli- ja võistluskirjandites 5 (*ja*, *et, kui, ning, aga*, koolikirjandites sidesõna *aga* asemel sagedasem sidesõna *kuid*). Eksamikirjandites oli sidesõnu 4 (*ja, et, kui, ning*), kirjakeele sagedussõnastiku põhjal kuulub kirjakeele 20 sagedaima lemma hulka 5 sidesõna (*ja, et, kui, aga, ning*). Kõikides allkorpustes esinesid ka siduvad asesõnad *kes* ja *mis*.

6.2.3.6. Verbid

Kõikides vaadeldavates allkorpustes on 20 sagedama lemma seas vähemalt 3 verbi, võistluskirjandite tekstides sarnaselt eesti kirjakeele sagedussõnastikuga 4 verbi. Kõikide vaadeldavate allkorpuste sagedaim verb ning sagedaim lemma üldse oli *olema*. Kõikides allkorpustes esines 20 sagedaima lemma seas ka verb *saama*.

6.2.3.7. Nimisõnad

Kirjandite lemmad erinevad kirjakeele sagedussõnastiku lemmadest selle poolest, et kõikides kirjandite alaliikides on esimese 20 sagedaima sõna seas ka nimisõnu – põhikoolitekstides *inimene*, teistes kirjandikogudes *inimene* ja *elu*.

Põhikoolitekstid sageduselt 12. lemma *noor* on tõlgendatav nii nimisõna kui omadussõnana, kuna lemmatiseerimine on tehtud automaatselt ning morfoloogiline ühestaja ei suuda sõna *noor* puhul sõnaliiki eristada.

6.2.3.8. Hulgasõnad

Peaaegu kõikides koolikirjandite kogudes (v.a võistluskirjandid) kuulus 20 sagedaima sõna sekka ka kvantorsõna. Põhikoolis oli selleks sõna *palju*, gümnaasiumi koolikirjandites sõnad *palju* ja *kõik* ning eksamikirjandites ja kirjandite koondis eristus hulgasõna *kõik*. Võib oletada, et õppijad kasutavad hulgasõnu siis, kui soovivad midagi

üldistada. Kirjakeele sagedussõnastikus kvantorsõnu 20 sagedaima lemma hulka ei kuulu.

6.2.3.9. *ei* ja *ka*

Kõikides uuritavates allkorpustes oli 20 sagedaima seas esindatud ka eitussõna *ei* ning rõhusõna *ka*.

Tabel 12. Emakeeleõppijakorpuse kahekümne sagedama lemma analüüs sõnaliigiti

Kasutatud tähistused:

X	**verb**
X	*asesõna*
X	nimisõna
X	***sidesõna, sidend***
X	*hulgasõna*

põhikooli-kirjandid	gümnaasiumi koolikirjandid	gümnaasiumi võistlus-kirjandid	gümnaasiumi eksami-kirjandid	õpilas-kirjandite koond	eesti kirjakeele sagedussõnastik
695 olema	3671 olema	11336 olema	1125 olema	15702 olema	olema
446 *ja*	1939 *see*	5899 *see*	598 *ja*	8181 *see*	*ja*
374 *mina*	1744 *ja*	5729 *mina*	541 *see*	7770 *ja*	*tema*
343 *see*	1009 *et*	5580 *ja*	277 *et*	6874 *mina*	*see*
295 *tema*	948 ei	3331 ei	272 *tema*	4484 *et*	*mina*
171 *et*	817 *tema*	3304 *et*	262 *mina*	4443 ei	ei
164 ei	771 *mina*	2507 *mis*	259 ei	3356 *mis*	*et*
143 *kui*	770 ka	2189 *kui*	254 *mis*	3191 *tema*	*kui*
122 *mis*	767 inimene	2079 *tema*	211 inimene	3053 *kui*	*mis*
116 ka	727 *mis*	2055 *oma*	204 *kui*	2803 ka	ka
102 *oma*	721 *kui*	1917 ka	204 ka	2680 *oma*	saama
101 noor/	678 *ning*	1857 *ise*	187 *oma*	2455 inimene	*oma*
99 saama	523 oma	1792 saama	149 *ise*	2446 *ning*	*aga*
96 minema	507 *ise*	1684 *ning*	140 elu	2445 *ise*	*sina*
86 *kes*	474 saama	1612 inimene	139 *ning*	2365 saama	*ise*
84 *palju*	413 *kes*	1422 elu	116 *kes*	1732 *kes*	*ning*
84 *ning*	376 *palju*	1233 *kes*	113 *nii*	1732 elu	*kes*
81 *ise*	366 *kõik*	1174 tegema	111 võima	1524 tegema	tulema
76 inimene	330 võima	1126 *aga*	111 *kõik*	1509 *aga*	pidama
76 *aga*	325 *kuid*	1103 pidama	107 saama	1508 *kõik*	*üks*

6.2.3.10. Sagedusanalüüs sõnaliigiti

Järgnevalt analüüsitakse asesõnu ja kvantorsõnu, verbe ning nimisõnu põhjalikumalt. Lisaks neile esitatekse iga allkogu kolm sagedamana esile kerkinud omadussõna ning kolm sagedamat pärisnime.

6.2.3.10.1. Tuumsõnade hulka kuuluvad sagedased pronoomenid ja kogusesõnad emakeeleõppija korpuses

Eesti keele tuumsõnade hulka kuuluvad pronoomenitest

a) personaalpronoomenid: *mina, sina, tema*;

b) demonstratiivid: *see, nii, selline*;

c) refleksiivpronoomenid: *oma, ise*;

d) indefiniitpronoomenid: *miski, keegi, mingi*. (Pajusalu *et al.* 2004: 12)

ja kogusesõnadest

a) kvantorid: *kõik, veel, enam, ainult, pool, palju, mõni, väga, kogu, iga*;

b) arvsõnad: *üks, teine, kaks, esimene*. (Pajusalu *et al.* 2004: 12)

Ilukirjandustekstides kasutatakse asesõnu oluliselt rohkem kui ajakirjandustekstides (Hennoste, Muischnek 2000: 205).

Õppijatekstides esinevate pronoomenite ja kogusesõnade analüüsimiseks on lemmatiseeritud sõnasagedusloendist välja kirjutatud 10 enamesinevat eesti keele tuumsõnade hulka kuuluvat asesõna ning kvantorsõna. Sõnad on esitatud allkorpuste kaupa nende esinemise järjekorras. (vt tabel 11).

Pronoomenite ning kogusesõnade kasutussagedus ei erine olulisel määral kirjakeele sagedussõnastiku andmetest: kolm sagedaimat pronoomenit on kõikides analüüsitud allkogudes *see*, *mina* ja *tema*; viienda pronoomenina esineb kõikides allkogudes *oma*. Õpilastel on esikohal pronoomen *see*, kirjakeeles *tema*.

Pronoomen *see* viitab lähimas kontekstis mainitud objektile, situatsioonile või situatsioonide kompleksile (ainsuses). *See* võib viidata ka elusolendile. Asesõna *see* kasutatakse palju esseedes ning ilukirjanduses (Meier 2003: 131).

Esile võib tuua, et kirjakeele sagedussõnastikuga võrreldes puudub õpilastekstide sagedamate asesõnade seast pronoomen *sina* (v.a põhikooliõpilaste tekstid). See on

seletatav nõudega vältida kirjandi kirjutamisel argist *sina*-stiili. Pronoomen *ise* jäi kõikides õpilastekstides esimese viie sagedama hulka, kuid kirjakeele sagedamate hulgas esile tõusnud ei ole. Lisaks loetletutele kuulub sagedamate pronoomenite sekka ka ebamäärane – *miski*.

Õpilastekstidele on iseloomulik, et ümbritsevat analüüsitakse *mina*-vaatepunktist, seetõttu on pronoomeni *mina* kasutussagedus pronoomeni *tema* kasutusest sagedam. Samuti viitavad enese isikule pronoomenid *oma* ja *ise.*

Kvantorsõnade esitamine koos asesõnadega näitab, et viie sagedama sekka koonduvad asesõnad, alates kuuendast sõnast tõusevad esile kvantorsõnad. Õpilastekstides on sagedad *kõik*, ja *palju*, *üks, teine* ja *iga* kirjakeeles aga *üks*, *kõik*, *teine* ja *kaks*.

Tabel 13. Emakeeleõppija korpuse tekstides esinenud sagedasemad asesõnad ning kogusesõnad.

	põhikooli-kirjandid	gümnaasiumi koolikirjandid	võistlus-kirjandid	eksami-kirjandid	õpilas-kirjandite koond	kirjakeele sagedus-sõnastik
1	mina	see	see	see	see	tema
2	see	tema	mina	tema	mina	see
3	tema	mina	tema	mina	tema	mina
4	oma	oma	oma	oma	oma	oma
5	**palju**	ise	ise	ise	ise	sina
6	ise	**palju**	**kõik**	**kõik**	**kõik**	**üks**
7	nii	**kõik**	nii	**üks**	**palju**	**kõik**
8	**kõik**	**teine**	miski	**teine**	nii	**teine**
9	**üks**	**üks**	**üks**	**iga**	**üks**	miski
10	sina	**iga**	**teine**	miski	miski	**kaks**

6.2.3.10.2. Emakeeleõppija korpuse sagedamad verbid (vt tabel 14)

Verbidest on eesti keele põhisõnavara operaatorite ehk tuumsõnade hulka arvatud (Pajusalu *et al.* 2004: 11)

1) üldverbid: *olema, tegema*

2) liikumis- ja asendiverbid: *tulema, minema, käima seisma, istuma*;

3) omandamisverbid *saama, võtma, andma, panema, viima, tooma;*

4) modaalverbid: *pidama, võima* (*saama*)

5) algusverb *hakkama*

6) kognitiivverbid: *nägema, tahtma, mõtlema, teadma, vaatama, kuulama, tundma, arvama*

7) suhtlusverbid: *ütlema, rääkima, küsima, vastama*

8) kausaalverbid: *ajama, laskma, jääma*

Ajakirjanduses on sagedaimad verbid: *olema, saama, pidama, tulema, jääma, hakkama, tegema, minema, ütlema, võima ja andma, tahtma, teadma, arvama.* Ilukirjandus erineb ajakirjandusest meeltetegevust kasutavate verbide poolest: verbid *nägema, vaatama, mõtlema, tundma.* (Hennoste, Muischnek 2000: 204)

Õppijatekstides esinevate verbide analüüsimiseks on lemmatiseeritud sõnasagedusloendist välja kirjutatud 20 enamesinevat verbi. Sagedamad verbid on esitatud allkorpuste kaupa nende esinemise järjekorras, st allkorpuses sagedaim verb on esitatud tabelis esimesena (vt tabel 14).

Igas allkorpuses on rasvase kirjaga ning kaldkirjas tähistatud need verbid, mis erinevad kirjandite koondi verbidest. Kirjandite koondis on rasvase kirjaga märgitud verbid, mis erinevad kirjakeele sagedussõnastiku esimese kahekümne verbi hulka kuuluvatest sõnadest. Kirjakeele sagedussõnastiku viis rasvase kirjaga märgitud verbi puuduvad kirjandite koondi 20 sagedaima verbi hulgast (kuigi võivad esineda kirjandite allkorpuses sagedaimate verbide seas).

Õppijatekstide koondi 20 sagedama verbi hulka kuuluvad ootuspäraselt üldverbid *olema* (1) ja *tegema* (3.)*;* omandamisverbid *saama* (2.), *andma* (20.), *võtma* (17.); modaalverbid *pidama* (4.), *võima* (5.), (*saama* (2.)); liikumis- ja asendiverbid *tulema* (6.), *minema* (8.), kognitiivverbid *tahtma* (7.), *teadma* (10.), *mõtlema* (13.); algusverb *hakkama* (9); kausaalverb *jääma* (11.) ja suhtlusverb *ütlema* (19.).

Õppijatekstide 20 sagedaima verbi hulka kuulub ka viis verbi, mida kirjakeele sagedussõnastiku sagedaimate hulgas ei ole: *õppima* (14.), *elama* (12.), *nägema* (15.), *arvama* (16.), *leidma* (18.).

Kognitiivverbid *nägema* ja *arvama* kuuluvad eesti tuumsõnade hulka. Verb *arvama* on T. Hennoste ja K. Muischneki analüüsi põhjal (2000: 204) 1930. ja 1990. aastate ajakirjandustekstides sagedaimate verbide ning *nägema* ilukirjandustekstide sagedaimate verbide seas. Õppijatekstidele omased verbid *õppima* ja *elama* ei kuulu

eesti keele tuumsõnade hulka ega ajakirjandus- ja või ilukirjandustekstide sagedaimate verbide hulka. 10 000 sagedaima eesti kirjakeele sõna seas on *õppima* alles 376. kohal (esines ajakirjandustekstides 168 korral, ilukirjandustekstides 138 korral, kokku 306 korda (Kaalep, Muischnek 2002: 158), kuid selle verbi kuulumine õpilastekstide 10 sagedaima verbi sekka ootuspärane, kuna peegeldab koolilaste põhitegevust. Verb *elama* on Eesti kirjakeele sagedussõnastiku põhjal 10 000 sagedaima sõna seas 125. (esines ajakirjandustekstides 288 ja ilukirjandustekstides 527 korda. (Kaalep, Muischnek 2002: 155).

Kirjakeele sagedussõnastiku 20 sagedaima verbi seas leidunud *jõudma* ja *käima* ning eesti keele tuumsõnad *rääkima* ja *panema* puudusid õppijate koondi tekstide 20 sagedama seast.

Järgnevalt analüüsitakse õppijakirjandite allkogudes 20 sagedamini esinenud verbidest neid, mis koolikirjandite koondanalüüsis sagedamate hulgast puudusid.

Enim üllatasid põhikoolikirjandid. Ootuspäraselt sagedate verbide hulgas (*olema, saama, minema, tegema, pidama, tulema, võima, hakkama, tundma, tahtma, nägema, arvama, teadma*) on põhikooliõpilaste tekstides enim selliseid verbe, mis teistes õppijatekstide allkorpustes sagedamatena silma ei paista, kuid on esindatud kirjakeele sagedussõnastikus. Kõik viis verbi, mis küll ei kajastunud kirjandite koondis, aga on kirjakeele sagedussõnastiku sagedamate seas, esinesid ka põhikooliõpilaste sagedamate hulgas. Nendeks on verbid *käima,* v*aatama, jõudma, rääkima* ja *panema.*

Miks sarnanevad just põhikooliõpilaste tekstid verbide valiku poolest enim kirjakeele sagedamate verbidega, vajab täpsemat uurimist. Ehk võib põhjus olla tekstide jutustavamas laadis, kuid see jääb praegu ainult oletuseks.

Põhikooliõpilaste sagedamate verbide sekka kuulunud *meeldima* ja *huvitama*, mis ei kuulu ka kirjakeelekorpuse sagedamate verbide hulka, tuumsõnade hulka, ajakirjandus- ega ilukirjandustekstides kasutatavate sagedamate verbide hulka, peegeldavad põhikooliõpilase suhestumist maailmaga ning on otseselt seotud õpilaste huvide ja hobide ehk vaba aja teemaga, kuigi käsitletud kirjandite teemad ei eelda, et kirjutised peaksid peegeldama õppijate vaba aja teemasid. Verb *meeldima* on eesti kirjakeeles 10 000 sõna seas sageduselt 404. (282 esinemust) ning *huvitama* 630. sõna (181 esinemust).

Koolikirjandite sagedamate verbide seas erinevad kirjandite koondis loetletud verbidest *muutma ja* tuumsõnade hulka kuuluv *tooma.* Verb *tooma* esines kirjakeele sagedusanalüüsi aluseks võetud ilu- ja ajakirjandustekstides 834 korda (ajakirjandustekstides 446 ja ilukirjandustekstides 388 korda), verb *muutma* 317 korda (224 ajak/ 93 ilukirj).

Võistluskirjandite koguhulk oli teistest kirjandite allkogudest mahu poolest suurem, mistõttu kõik võistluskirjandi verbid kajastusid ka kirjandite koondis.

Eksamikirjandite sagedamad verbid, mida õppijate koondtekstide sagedamate verbide hulgas ei leidu, on *jõudma*, *tundma* ja *tunduma*, Verb *jõudma* puudus teiste gümnaasiumiealiste õppijate allkogude sagedamate verbide seast, kuid on kirjakeele sagedussõnastiku 20 sagedama verbi hulgas 17. Kirjakeele sagedussõnastiku andmetel esineb verb *tundma* 10 000 sagedama kirjakeele sõna seas 1205 korda ning on iseloomulik pigem ilukirjandustekstidele (335 esinemist ajakirandustekstides ja 870 esinemist ilukirjandustekstides). Verb *tunduma* on kirjakeeles sageduselt 224. lemma (497 esinemust/ajakirjandustekstides 160/ ilukirjandustekstides 337).

Õppijatekstide sagedamate verbide üldisemaks iseloomustamiseks võib öelda, et tavapäraste tuumsõnade kõrval olid õppijatekstides esindatud ka harvemad verbid nagu *elama* (12. verb), *õppima* (14. verb), *nägema* (15.), *arvama* (16.), *leidma* (18.) ning allkorpustes *meeldima* (põhikooliõpilaste tekstides), *huvitama* (põhikooliõpilaste tekstides), *muutma* (gümnaasiumiõpilaste koolikirjandites), *tooma* (gümnaasiumiõpilaste koolikirjandites), *tunduma* (gümnaasiumiõpilaste eksamikirjandites) ja *tundma* (gümnaasiumiõpilaste eksamikirjandites).

Sagedate verbide konkordantse

põhikooliõpilaste sage verb ***käima***	koolikirjandite sage verb ***muutma***	eksamikirjandite sage verb ***jõudma***
Ma *käisin* juulikuus haapsalus poolsaarel *käisime* me delfinaari mured *käivad* kaasas iga teisme mured *käivad* inimesega alati veel *käisin* ma esmaabilaagris see *käib* ka veel paljude teiste sportida, *käia* kinos, õppida karjas *käima* – suvel küll osad mõnikord *käisid* nad streetrace´il Nad *käisid* sageli ka 29 Võrus Päeval *käisid* isa ja vennad tööl	on *muutnud* olukorda maal seda *muuta* saab vaid parandamiseks *muuta*. Ajaloo paemaks *muuta*, põhjuseks *muuta*, investeerida tulevikku vajadust *muuta* raiskavaid tarbi seda *muuta*. Kas siis kardavad maailma *muuta* : haridust väärtusta *muuta*, lastes oma poja hukata saab *muuta*. nendele küsimustele võimalik *muuta*, ei tea keegi	mõningad *jõudmas* kodanlikku sinna *jõuda*. Ajakirjandus oli kaugemale *jõudnud*, poeb üks on *jõudnud* juba päris suurte *jõudnud*, mis tuli küll nii kuhugi *jõuda*, peab selleks ka et *jõuda* sinna, kuhu vaja elus *jõuda* võimalikult kõrgele puhastatuna *jõuda* maise elu et *jõuda* sihtkohani.

Tabel 14. Emakeeleõppijate tekstide sagedamad verbid võrrelduna kirjakeele korpuse sagedamate verbidega.

	põhikooli-kirjandid	gümnaasiumi koolikirjandid	gümnaasiumi võistlus-kirjandid	gümnaasiumi eksami-kirjandid	õpilas-kirjandite koond	kirjakeele sagedus-sõnastik
1.	olema	olema	olema	olema	olema	olema
2.	saama	saama	saama	võima	saama	saama
3.	minema	võima	tegema	saama	tegema	tulema
4.	tulema	pidama	pidama	pidama	pidama	pidama
5.	tegema	tegema	tulema	tulema	võima	ütlema
6.	hakkama	tulema	võima	tegema	tulema	tegema
7.	pidama	elama	tahtma	tahtma	tahtma	minema
8.	võima	minema	minema	minema	minema	võima
9.	nägema	jääma	hakkama	jääma	hakkama	jääma
10.	***käima***	võtma	teadma	arvama	teadma	teadma
11.	***meeldima***	teadma	õppima	andma	jääma	võtma
12.	***huvitama***	hakkama	mõtlema	elama	**elama**	hakkama
13.	ütlema	nägema	jääma	teadma	mõtlema	andma
14.	võtma	tahtma	elama	leidma	**õppima**	tahtma
15.	tahtma	***muutma***	arvama	võtma	**nägema**	**vaatama**
16.	***vaatama***	arvama	nägema	***jõudma***	**arvama**	**rääkima**
17.	***rääkima***	mõtlema	leidma	hakkama	võtma	**panema**
18.	***panema***	andma	ütlema	ütlema	**leidma**	**jõudma**
19.	***jõudma***	***tooma***	võtma	***tundma***	ütlema	**käima**
20.	teadma	ütlema	andma	***tunduma***	andma	mõtlema

6.2.3.10.3. Emakeeleõppija korpuse sagedamad nimisõnad (vt tabel 15)

Sõnavara on keele kõige liikuvam ja muutlikum osa. Kui muutub meid ümbritsev tegelikkus, siis muutuvad ka sõnad, millega me oma tegudest ja mõtetest räägime. (EKK 1997: 475) Nimisõnu võib nimetada teksti sisusõnadeks. Sisusõnad kajastavad kirjutaja maailmapilti ning on seotud teemaga, millest kirjutatakse.

Järgnevalt analüüsitakse sagedamaid nimisõnu allkorpuste kaupa. Tabelis 15 on kirjandite allkorpustes kaldkirjaga tähistatud need sõnad, mida kirjandite koondis ei esinenud. Kirjandite koondis on rasvase kirjaga markeeritud sõnad, mis erinevad kirjakeele sagedamatest nimisõnadest ja kirjakeele sagedamate seas on tähistatud need, mis erinevad emakeeleõppija koondtekstide sagedamatest.

Huvipakkuvad tundusid ka sagedamad omadussõnad ning õppijakeele allkorpustes enamesinenud pärisnimed, mistõttu esitatakse iga uuritava allkorpuse kohta ka need.

Põhikooliõpilaste kirjandid

Põhikooliõpilaste kirjandite sisusõnad on põhjendatavad õpilast ümbritseva igapäevase keskkonna järgi. Põhikooli lapse maailma kuuluvad tema *klass*, kus on kõige olulisemaks *sõber*, olgu siis *tüdruk* või *poiss*. Teine oluline koht on *kodu* ning lähedastest inimestest eristub *vend*. Põhikooliõpilase ajaarvamine käib *aasta* ja teistest allkorpuse sagedustest erinevalt *päeva* kaupa. Loomulikult räägitakse kirjandites oma *mure*dest, milleks sagedussõnastiku põhjal otsustades on sageli *vägivald*. Põhikooliõpilase jaoks on olulised tema *hobid* sh *muusika*.

Üllatavate sõnadena esinevad sageli *kord, küülik* ning *mootor*. Sageduses selgusele jõudmiseks analüüsiti esmalt kirjandite pealkirju, kuid nende sõnade sagedus pealkirjade järgi eeldatav ei ole. Kollokatsioonide analüüs näitab *musta mootori* ja *kass Mootori* kasutamist. Nimisõna *küülik* esineb sageli ühes tekstis. Sõna *kord* esineb muinasjutule tüüpilise algusena *elas kord* (*Elasid kord 3 poissi / Elas kord üks pisike hiiretüdruk / Elas kord kolm noort ja ilusat*).

Ühes (mõnes) kirjandis sageli korratavad sõnad ei iseloomusta põhikooliealiste laste tekste üldiselt ning seetõttu on sagedusloendis nende kõrvale kirjutatud ka teine sõna, et soovi korral oleks võimalik juhuslikult sagedana esiletõusnud sõnad tähelepanuta jätta.

Omadussõnadest on põhikooliõpilaste kirjandites sagedaimad üldised omadussõnad (*noor*), *suur* ja *hea*. Pärisnimedest eristusid enimkasutatavatena: *Mary, Ray, Minni.*

Gümnaasiumiõpilaste koolikirjandid

Koolikirjandite allkorpuses eristusid sagedamate nimisõnadena *naine, mees, riik, saatus, probleem* ja *isa.* Põhikoolis õppijatega võrreldes on vaateväli muutunud avaramaks, sest *klassi* asemel räägitakse *Eestist*, *riigist* ning *maailmast* üldisemalt. *Poisi* ja *tüdruku* asemel räägitakse *mehest* ning *naisest* ja lihtsakoelisest *murest* on saanud *probleem*. Koduga seotud inimestest esineb tekstides rohkelt *isa* kuju, kuid kas positiivse või negatiivse tegelasena, ei ole võimalik kindlalt öelda. Märgatavalt on langenud sõna *noor* kasutamissagedus. Sageli arutletakse hoopis *saatuse*, *töö, raha* ning *võimaluste* üle.

Omadussõnadest olid sagedaimad *suur* ja *hea*, pärisnimedest *Eesti*, *Euroopa* ja *Hamlet*.

Gümnaasiumiõpilaste võistluskirjandid

Võistluskirjandid käsitlevad paljuski karjääriteemat ning seetõttu lisanduvad koolikirjandite sisusõnadele *karjäär, unistused, eesmärgid* ja *tulevik*. Eristuvaks sõnaks on ka *mõte.* Uuesti on tähelepanu pööratud *klassile,* mille *valik* võib määrata edasise *tee* ja *tuleviku.*

Omadussõnadest olid sagedaimad taas *hea* ja *suur* ning pärisnimedest *Eesti, Tartu, Tallinn* ja isikunimi *Toomas*.

Gümnaasiumiõpilaste eksamikirjandid

Eksamikirjandites üritavad gümnaasiumiõpilased näidata end küpse kodaniku ning erudeeritud inimesena. Räägitakse *kunstist* ning *ajakirjandusest*; probleemid on kuhjunud *kriisiks*, aruteltakse *ühiskonna*, *riigi* ja *rahva* üle.

Sisusõnade kohta võib kokkuvõtlikult öelda, et 20 kirjandite koondis esinevat sagedamat sisusõna, mis kirjakeele sagedamatest eristuvad, peegeldavad hästi noorte mõttemaailma: *tulevik, noor, Eesti, võimalus, karjäär, maailm, unistus, õnn, kool, tänapäev, eesmärk, riik.*

Omadussõnadest eristusid sagedaimatena *suur* ning asesõna *selline*, pärisnimedest olid sagedaimad *Eesti, Faust* ja *Tammsaare*.

Sagedaid nimisõna konkordantse õppijakirjandites

sõber	***kunst***	***probleem***
head *sõbrad* ja pole vaja tülid, *sõbra* reetlikus, õnnetu minu *sõber* ühel hetkel rääkis niisugused *sõbrad*, keda tunne elurõõmsad *sõbrad*. Nendega toredad *sõbrad* ning mõnus aja peab *sõber* olema aus ja sõbralik oma *sõbraga* läksin jalgratastega mu *sõbra* viia politseisse. Kuid	on *kunst* jääv meie ühiskonnas kindlasti *kunst*. Oma maailmapil on *kunst* vaid olemasolev ootam etendab *kunst* selles elutähtsat loovad *kunsti*, ning teised, kes sest *kunsti* mõistmine eeldab kuid *kunst* ei kaota kunagi jääb *kunstile* igaveseks kindel *kunsti* poolt seletamatu saab aga *kunst* liigub pidevalt	tähtsamaks *probleemiks* on see *probleem* endiselt lahendami võrdõiguslikkuse *probleem* ei *probleeme* erinevate haiguste sellel *probleemil* kaks poolt teeb *probleemist* tegeliku või suur *probleem* tehtud tarbijakaitse *probleeme* eemal *probleemide* keerisest isiklikele *probleemidele*

Tabel 15. Sagedamad nimisõnad emakeeleõppijate kirjandites võrrelduna kirjakeele sagedussõnastiku sagedamate nimisõnadega.

	põhikooli-kirjandid	gümnaasiumi koolikirjandid	gümnaasiumi võistlus-kirjandid	gümnaasiumi eksami-kirjandid	õpilas-kirjandite koond	kirjakeele sagedus-sõnastik
1.	noor	inimene	inimene	inimene	inimene	aasta
2.	inimene	elu	elu	elu	elu	**mees**
3.	*sõber*	eesti	tulevik	ajakirjandus	**tulevik**	aeg
4.	aeg	aeg	töö	ühiskond	töö	inimene
5.	tänapäev	*naine*	aeg	maailm	aeg	**naine**
6.	elu/ *küülik*	aasta	karjäär/aasta	aeg	aasta	**käsi**
7.	*kord*/asi	riik	unistus	riik	**noor**	**päev**
8.	*mure*	*saatus*	noor	*kunst*	**eesti**	**kroon**
9.	*huvi*	*mees*	võimalus	*naine*	**võimalus**	laps
10.	*vend*	laps	maailm	*mees*	laps	asi
11.	laps	maailm	õnn	eesti	**karjäär**	**pea**
12.	*klass*	raha	asi	*tee*	asi	**riik**
13.	aasta	asi	laps	mõte	**maailm**	**silm**
14	*tüdruk*	tänapäev	raha	aasta	**unistus**	töö
15.	*poiss*	töö	eesmärk	*rahvas*	**õnn**	elu
16.	*kodu*	õnn	*mõte*	laps	raha	raha
17.	*mootor/ vägivald*	noor	eesti	eesmärk	**kool**	**linn**
18.	*päev*	*probleem*	kool	*areng*	**tänapäev**	**miljon**
19.	kool	võimalus	*valik*	kriis	**eesmärk**	**ema**
20.	*muusika*	*isa*	*tee*	asi	**riik**	**maja**

Emakeeleõppijate statistiline analüüs kinnitas, et gümnaasiumiõpilase essee keskmine lausepikkus vastab 20. sajandi kirjakeele tekstide keskmisele lausepikkusele. Kirjand sisaldab keskmiselt 578 sõna ja on u 44 lause pikkune, põhikooliõpilastel vastavalt 282 / 26.

Korpuse analüüsi põhjal võib väita, et õppijad kasutavad nimisõnu keskmisest rohkem. 20 sagedama nimisõna seas olid sagedaimad *inimene* ja *elu*, mis eristusid ka lemmade analüüsis sagedatena.

Õpilastekstide sagedamateks pronoomeniteks on *oma* ja *ma* ning kvantorsõnadeks üldistavad *kõik* ja *palju*. Verbikasutuselt erinevad õpilastekstid kirjakeelest verbide *elama*, *õppima*, *nägema, arvama, leidma* kasutussageduse poolest.

Statistilise analüüsi kokkuvõtteks tasub meenutada, et 20 esimest lemmat katavad tekstist u 24,6% (Kaalep, Muischnek 2002: 14).

6.3. Emakeeleõppija korpuse veakategooriad ja märgendussüsteem

Märgendussüsteemi loomisel lähtun S. Grangeri (2003) seisukohtadest õppijakorpuse vigade märgendamise kohta (vt pt 5.2.2.): korpuse annotatsioon peab olema informatiivne, korduvkasutatav, kohandatav, järjekindel, lähtuma õppijate tegelikust keelekasutusest ning olema märgendaja jaoks hallatav.

Õppijakorpuse veamärgendussüsteemi etappidest kirjeldatakse 1) veakategooriate loomist, veakategooriaid ning neisse kuuluvaid veatüüpe, 2) märgendussüsteeemi ja märgendamist, 3) veaparanduste sisestamist.

6.3.1. Vigade jagamine üldkategooriatesse

Keel kui süsteem eksisteerib keele kasutajate teadvuses keeleoskusena, st märkide ja reeglite kogumina, mida kasutades inimene on võimeline oma mõtteid teistele edasi andma (EKK 1997: 25). Vead keelekasutaja tekstis näitavad, et sõnumi edastamine ei ole korrektne, vaid millegi poolest puudulik või häiritud.

Eesti laste keelevigade leidmiseks alustati õppijate tekstide kogumisest ehk keelekorpuse koostamisest (vt pt. 6.2.1. ja 6.3.). Veataksonoomia loomiseks alustati tekstides leiduvate vigade tähistamisest ja parandamisest (vt pt 5.2.1.). Katmaks erineva kirjutamisoskusega õppijate tekste, valiti analüüsimiseks 10 000 sõna suurune hulk erineva pikkusega gümnaasiumiõpilaste kirjandeid Viljandi, Otepää ja Tartu linna ning maakonna õpilastelt. Kirjandi pikkust tõlgendati kui ühte õpilase keelelise võimekuse näitajat. Järgnes kirjandites leiduvate vigade tähistamine ja parandamine, mille eesmärgiks oli koondada vead kategooriatesse.

Õppijate tekstides on vead põimunud ning alati ei ole võimalik üheselt otsustada, missuguse veaga tegu on. Ühte veakategooriasse kuulumine ei välista keelendi kuulumist samal ajal ka teise kategooriasse. Näiteks lauses <KOK/GÜ/2007/VIL/21> *Juba Vladimir Nabokovi teoses "Lolita" on hea näide sellest, kudas inimene võib olla meediast sõltuv* esinevat sõna *kudas* võib liigitada vähemalt nelja veatüüpi:

1) stiiliviga – argikeelsus
2) näpuviga – teksti sisestamisel tehtud viga
3) ortograafiaviga
4) olematu sõna valimine

Selleks, et märgendatud tõlgendused oleksid analüüsitavad, koostati lihtne ja paindlik kooligrammatika käsitlusest lähtuv klassifikatsioon koos õpilastekstidest võetud näidetega.

Veakategooriate aluseks võeti hierarhilise keelesüsteemi tasandid. Kirjaliku teksti keeletasanditele – morfoloogiale, süntaksile, leksikale ja tekstile (EKG: 326) lisanduvad veel rakendusliku suunaga stilistika ja ortograafia (tähenduses õigekirjutus ehk õpetus kirjakeele õigeist kasutusnormidest (EKK 1997: 27)). Sõnamoodustuse tasand jaotati kahe veakategooria vahel: õigekirjareeglite vastu eksimist käsitletakse ortograafia kategoorias, sisutähenduse vigu semantika kategoorias. Eesti keele hierarhiliste tasandite ja kooligrammatikas levinud käsitluse (Ehala (1998) ja Valmis, Valmis (2001) õpikute ning ning Ehala (1997) ja Ehala *et al.* (2005) metoodikaraamatute põhjal) võrdlemine ning õppijate tekstidest leitud vigade analüüsimine andis tulemuseks seitse üldist veakategooriat. Üldtasandil eristatakse ortograafiavigu, interpunktsioonivigu, morfoloogiavigu, süntaksivigu, semantilisi vigu, stiilivigu ja tekstivigu. Igaks juhuks on loodud ka üldkategooria nimetusega *muud vead*, mida tuleb kasutada siis, kui viga loetletud üldkategooriatesse ei liigitu.

6.3.2. Veakategooriad

Ortograafia tähistab veakategooriate jaotuses *õigekirjale* ehk *õigekirjutusele* kui kirjakeele normile mitte vastavate keelendite kategooriat. Ortograafia kategooriasse kuuluvad ortograafiliselt vigased, normist hälbivad keelendid. Õigekiri ehk *normitud kiri* (Ehala 2005: 34) hõlmab häälikuortograafiat, poolitamist, kokku- ja lahkukirjutamist ning algustähe ortograafiat.

Interpunktsioon ehk kirjavahemärgistus on ühel ajal nii normitud kirja kui ka süntaksi valdkond. Käesolevas jaotuses on peetud otstarbekaks interpunktsioonivead eraldi veakategooriasse liigitada. Interpunktsiooni veakategooriasse kuuluvad kõik kirjavahemärgistamise vead.

Morfoloogia ehk vormiõpetus on grammatika osa, mis tegeleb sõnavormidega – nende moodustamise ning nendest arusaamisega. Morfoloogia kategooriasse kuuluvad kõik vead, mis on seotud sõna tüve või sõnavormi moodustamisel tekkinud

kõrvalekalletega. Veakategooriate liigituses jagatakse morfoloogiavead omakorda sõnaliikide alusel.

Süntaks ehk lauseõpetus on keeleteaduse osa, mis tegeleb sõnadest ja sõnavormidest kõrgemale jäävate keelesüsteemi tasanditega, kirjeldab, missugused on nende osade seosed ja funktsioonid. Süntaksi veakategooriasse kuuluvad: rektsiooni, ühildumise ja sõnajärjevead, täiendi, lisandi ja lauselühendiga seotu, käände valimine ning ajaline, arvuline, kõneviisi ja tegumoe ühtsus. Süntaksivigade alla kuuluvad ka need laused, millest on mingi oluline lauseliige puudu.

Semantika ehk tähendusõpetuse kategooriasse kuuluvad sõnade (sõnasemantika) ja püsiühendite ehk fraseologismide tähendusega seotud vead. Leksika ning semantika kuuluvad veataksonoomias ühte kategooriasse, kuna need kaks on lahutamatult seotud: Sõna kasutamine vales tähenduses näitab, et õppija ei tea selle sõna tähendust või teab seda ebamääraselt.

Stiil on "suhtlussituatsioonile ja -eesmärgile vastav keelekasutusviis, mille tunnusteks on isepärasused morfoloogias, sõnamoodustuses, sõnavaras, lausestuses, kujundikasutuses, teksti liigenduses vm keele valdkondades". (EKK 1997: 508) Rakenduslik kategooria *stiil* on kirjandites täiesti eraldi käsitletav ning hõlmab stiilivariatsioone, mis on omased erinevatele registritele. "Eestigi keeles tunneme paljude sõnade ja väljendite juures selgesti stiilivärvingut, s.o ühe kindla stiili omadusi. Need keelendid on erilised, rõhutatud. /.../ markeerimata, s.o harilikud, tavapärased, tavakeskmised, standardsed keelendid sobivad oma neutraalsuses igale poole ega tekita kuskil tõrkeid" (EKK 1997: 508).

Stiilivea kategooriasse kuuluvad stiililiselt markeeritud sõnad ja stiililiselt mittesobivad (näiteks liiga pikad) laused. Kui eelmiste jaotuste puhul ei olegi alati oluline, millises registris keelt kasutatakse, grammatiliselt on õige ikka sama keelevariant, siis stiili õigsus ja väärus sõltub kontekstist. Stiili all eristatakse nö stampväljendite, moesõnade kasutamist, stiilivärvinguid, liiga pikki või liiga lühikesi hakitud lauseid.

Tekstivigade hulka liigitatakse viitamisega (deiktiliste sõnadega) seotud vead. Deiksise all mõistetakse nähtust, mis seob keele ja konteksti. Deiktilised sõnad seovad sõnu, mille kasutamine ja interpreteerimine sõltub situatsioonist, mis valitseb nende lausumise hetkel (Pajusalu 1999: 17). Diskursuse vigade kategooriasse kuuluvad peale valesti viitavate pronoomenite ning ebaselgelt viitavate sõnade ka sidususprobleemiga laused, kaheti mõistetavad laused, halva sõnastuse või mõistetamatu sisuga laused.

Muu kategooria sisaldab nähtusi, mis eelnimetatud kategooriate alla ei liigitu.

Kogutud vigu liigitades ei ole seni olnud vajadust kasutada 8. kategooriat *muu*, kuigi igaks juhuks on selline võimalus jäetud.

6.3.3. Viga ja veapiir emakeeleõppija korpuses

Koolis ning eksamikirjandites parandatakse vigu õppijale hinde või hinnangu andmiseks. Emakeeleõppija korpuses on vigade märgendamisel teistsugune eesmärk. Märgenduse alusel ei hinnata õppijat, vaid kogutakse teavet õppijate keelekasutuse, sh ka vigade kohta. Seetõttu tuleb õppijakorpuses üldkeelest erinev tähistada vastava markeeringuga. Soovitav on märgendada ka ebasoovitavad, kuigi lubatud keelendid. Näiteks soovitatakse ajakirjade pealkirju tähistada suurte algustähtedega, kuid topeltmarkeeringu – jutumärkide – lisamist ei peeta äärele märgitavaks veaks. Õppijakirjandite märgendamisel tuleb sellised juhtumid märgenditega tähistada. Märgendite süsteem on paindlik ning vajadusel võib tähistada ka näiteks piltlikud väljendid vm uurijale huvipakkuv. Stiili hindamisel tuleb lähtuda kirjandi enda kontekstist.

Veaanalüüsi täpsuse huvides peavad märgendatavad üksused olema võimalikult konkreetsete piiridega. Iga vea puhul tuleb otsustada vea ulatus ning tähistada nii selle algus kui ka lõpp vastava märgendiga. Veapiiri väikseimaks üksuseks on sõna. Vajadusel võib veana tähistada lause või vajadusel mitu lauset (näiteks hakitud stiili puhul). Ühe märgendiga ei ole soovitav tähistada mitmeid lõike. Korduva vea puhul tuleb teksti lisada uus märgend. Veakategooriaid tähistavate märgendite piirid võivad, kuid ei pruugi kattuda.

6.3.4. Vigade märgendamine

Vigade tähistamiseks on eeskuju võetud S. Grangeri (2003) prantsuskeelse õppijakorpuse FRIDA mitmetasandilisest märgendite süsteemist, mis on paindlik, võimaldab teha erineva ulatusega otsinguid ning vajadusel märgendeid lisada.

Iga teksti peab märgendama kaks sõltumatut parandajat, kelle märgendeid on võimalik võrrelda. Märgendada on võimalik kas veakategooriate kaupa või lausete kaupa. Märgendaja peab alustekstina kasutama alati ilma märgenditeta nö puhast teksti.

Erinevate tõlgenduste korral saavad osad sõnad mitu märgendit. Need on nn hägusad alad. Mitmekordse märgenduse saavad ka sõnad, mis kuuluvadki erinevatesse veakategooriatesse. Näiteks sõna või fraas, milles on samal ajal nii ortograafia-, stiili- kui ka sõnajärjeviga vm.

6.3.4.1. Märgendid

Õppijate vigade üldklassifikatsiooni tähisteks on valitud üksikud tähed, näiteks morfoloogiavigade puhul <M>, süntaksivigade puhul <X> jne (vt tabel 16). Veatüübi tähistamiseks tuleb igale veale lisada kaks kahepoolset tähist sel moel, et vigane vorm (või väljend) jääb algust < >< > ja lõppu </ ></ > tähistavate märgendite vahele. Märgendi ja keelendi vahele peab jääma tühik.

Näiteks käändsõna morfoloogiline viga

<M><KS> isasi </KS></M>

pöördsõna morfoloogiline viga

<M><PS> jookstakse </PS></M>

Märgendi mõlemad pooled koosnevad kahest osast seetõttu, et oleks võimalik teha erineva ulatusega otsinguid. Näiteks märgend <M><PS></PS></M> võimaldab analüüsida kas kõiki morfoloogia ülemkategooriasse kuuluvaid <M></M> vigu korraga (väiksemaid veatüüpe eristamata) või ainult ülemkategooria alla kuuluvaid pöördsõnavigu <M><PS></PS></M> teistest veatüüpidest eraldi. Kui eesmärgiks on üldisem uuring, võib vigu märgendada ka ainult ülemkategooriate kaupa. Vajadusel saab täpsustavaid märgendeid ka hiljem lisada.

Korpuse märgendite seas on kaks tähist, mille lisamisel ei jää märgendi osapoolte vahele midagi. Need on interpunktsioonivea tähistamiseks mõeldud märk <I><PMARK></PMARK></I>, mis kannab sisu: *siit on interpunktsioonimärk puudu* ja süntaksivea tähistamiseks mõeldud märk <X><PUUDU></PUUDU></X>, mis kannab sisu: *siit on sõna puudu*. Kahepoolset märki on vaja paranduste sisestamiseks.

Spetsiifilisemaks nö kolmanda tasandi analüüsimiseks on võimalik veatüübi märgendile (nii algust kui lõppu tähistavasse märgendi ossa) uute tähiste lisamisega selekteerida kõik uurijat huvitavad keelendid.

Kui samale keelendile on vaja lisada mitu märgendit, siis peab need lisama sõnale sümmeetriliselt ehk nii, et **välimine** märgend jääb välimiseks nii sõna ees – alustavana kui ka sõna järel – lõpetava märgendina ning **sisemine** märgend on sisemine nii vea ees kui ka järel.

<M><KS> roheliseim **</KS></M>**

Tabel 16. Üldised veakategooriad ja nende tähised emakeeleõppija korpuses

veakategooria	veakategooria tähis
ortograafia	<O></O>
interpunktsioon	<I></I>
morfoloogia	<M></M>
süntaks	<X></X>
semantika	<N></N>
stiil	<S></S>
diskursus	<D></D>
muu	<U></U>

Igasse üldkategooriasse kuulub mitu veatüüpi. Kategooriate ning veatüüpide eristamiseks esitatakse esmalt ülevaatlik tutvustus ning seejärel veatüübid koos õppijatöödest võetud näidetega.

Ortograafiavead <O></O>

Ortograafiavigade alla liigitatakse õigekirjavead, algustähevead, kokku- ja lahkukirjutamise vead. Eraldi alaliigiks on teksti sisestamisel tekkinud vead, tsiteeringu markeerimiseks kasutatavate jutumärkide puudumine.

ortograafia	<O></O>	täheortograafiaviga aps – teksti sisestamise viga algustäheortograafia viga sõna peab olema kirjutatud kokku sõna peab olema kirjutatud lahku muu	<O><OK></OK></O> <O><APS></APS></O> <O><AT></AT></O> <O><KOKKU></KOKKU></O> <O><LAHKU></LAHKU></O> <O><ORT></ORT></O>

Interpunktsioonivead <I></I>

Interpunktsioonivigadena tähistatakse puudu olevad kirjavahemärgid, üle olevad ehk liigsed kirjavahemärgid ning valed kirjavahemärgid.

inter-punktsioon	<I></I>	liigne kirjavahemärk puuduolev kirjavahemärk vale kirjavahemärk muu	<I><LMARK></LMARK></I> <I><PMARK/></PMARK></I> <I><VMARK></VMARK></I> <I><MARK></MARK></I>

Morfoloogiavead <M></M>

Morfoloogiavigade alla liigitatakse sõna moodustamisel tekkinud vead. Eristatakse käändsõna, pöördsõna ning asesõna moodustamisel tehtud vigu. Morfoloogiavead, mis selle liigituse alla ei mahu, kuuluvad alaliiki *muu* viga.

morfoloogia	<M></M>	käändsõnaviga pöördsõnaviga asesõnaviga muu	<M><KS></KS></M> <M><PS></PS></M> <M><AS></AS></M> <M><MORF></MORF></M>

Süntaksivead <X></X>

Süntaksivead on liigitatud väljajättevead, rektsioonivead, lauselühendi kasutamise vead, järeltäiendi vead, ühildumisvead, vale ajavormi kasutamise vead, vale arvu kasutamine, vale käände kasutamine, sõnajärjevead. Süntaksivead, mis eelpoolnimetatute alla ei kuulu, liigitatakse alaliiki *muu*. Vajadusel luuakse viimasesse liigitusse koondunud vigadest uus alaliik.

Süntaksivigadena võib käsitleda ka kordust, nominalisatsiooni ning teatud sõnade esinemisest tulenevaid konstruktsioonijärjendeid, mis antud käsitluses on liigitatud stiili alla, kuna kooligrammatikas käsitletakse seda tüüpi eksimusi pigem stiilivigadena.

süntaks	<X></X>	väljajätteline lause rektsioon lauselühend järeltäiendi viga ühildumisviga vale aeg vale arv vale tegumood vale kääne sõnajärje viga muu	<X><PUUDU></PUUDU></X> <X><REK></REK></X> <X><LL></LL></X> <X><JT></JT></X> <X><UH></UH></X> <X><VT></VT></X> <X><VA></VA></X> <X><VTM></VTM></X> <X><VK></VK></X> <X><SNJRG></SNJRG></X> <X><SNX></SNX></X>

Semantikavead <N></N>

Semantikavigade hulka on liigitatud sõnavara piiratusest või sõnade tähenduse mittetundmisest tingitud vead tähendusvead ja sõnavalikuvead, parafraseerimine, tähenduse topeldamine, fraseologismivead ning reformuleerimisest tulenevad vead ning mitme fraseologismi segamisest tingitud vead.

Semantika	<N></N>	tähendus- või sõnavalikuviga parafraseerimine topeldamine fraseologismiviga reformuleerimisest tulenev segunemine ehk sulam muu	<N><TAH></TAH></N> <N><PF></PF></N> <N><TP></TP></N> <N><FR></FR></N> <N><RF></RF></N> <N><SEM></SEM></N>

Stiilivead <S></S>

Stiilivigade alla on liigitatud stiilivärvinguga sõnade kasutamine (argikeelsed sõnad), koolikirjandites taunitav argikeelne *sina*-stiil, moesõnade kasutamine, nominalisatsioon, kordus s.h ka grammatiliste vormide kordamine, liigne sõna lauses ning teksti lugemist segavad hakitud või liiga pikad laused, tuleviku väljendamise vead.

stiil	<S></S>	argikeelne sõna või väljend sina-stiil moesõna kasutamine nominalisatsioon kordus grammatiliste vormide kordamine liigne sõna, fraas lauses hakitud stiil liiga pikk lause tuleviku väljendamise viga muu	<S><ARGI></ARGI></S> <S><SINA></SINA></S> <S><MOE></MOE></S> <S><NOM></NOM></S> <S><KORD></KORD></S> <S><GKORD></GKORD></S> <S><LIIG></LIIG></S> <S><HAK></HAK></S> <S><PIK></PIK></S> <S><TUL></TUL></S> <S><STIL></STIL></S>

Tekstivead <T></T>

Tekstivigade alla liigitatakse viitamisega seotud vead – vale asesõna kasutamine ebamääraselt viitavad sõnad ning sisuvead.

Tekst	<T><T/>	vale asesõna kasutamine ebaselgelt viitav sõna kahetimõistetavus sidususviga sõnastusviga sisuviga muu	<T><VASES></VASES></T> <T><?VIIT></?VIIT></T> <T><KAH></KAH></T> <T><SID></SID></T> <T><SÕN></SÕN></T> <T><SISUL></SISUL></T> <T><TXT></TXT></T>

Muud vead <U></U>

Kui ülemklassi *muu* hulka koguneb palju ühte tüüpi vigu, on võimalik tekitada uus veakategooria.

muud	<U>	määratlemata vead	<U><MUU></MUU>

6.3.4.2. Märgendamine

Märgendamisel lisatakse veale alltüübi märgend.

Näiteks õppijakeele lause

<KOK/GÜ/2008/OTP/106> On olemas ütlus, viimane piisk karikast ja mindi üle piiri.

Vead ja nende markeerimine

süntaks
arvuline ühildumatus
<KOK/GÜ/2008/OTP/106> On olemas <X><VA> ütlus </VA></X> ,viimane piisk karikast ja mindi üle piiri.

vale käände kasutamine

<KOK/GÜ/2008/OTP/106> On olemas ütlus,viimane piisk <X><VK> karikast </VK></X> ja mindi üle piiri.

interpunktsiooniviga – vale kirjavahemärgi kasutamine

<KOK/GÜ/2008/OTP/106> On olemas ütlus <I><VMARK> , </VMARK></I> viimane piisk karikast ja mindi üle piiri.

ortograafiaviga – tsiteeringu õigekiri

<KOK/GÜ/2008/OTP/106> On olemas ütlus, <O><OK> viimane piisk karikast </OK></O> ja <O><OK> mindi üle piiri </OK></O>.

semantiline viga – fraseologismi mittetundmine

<KOK/GÜ/2008/OTP/106> On olemas ütlus, <N><FR> viimane piisk karikast </FR></N> ja mindi üle piiri.

semantiline viga – sõnavalikuviga

<KOK/GÜ/2008/OTP/106> <N><TAH> On olemas </TAH></N> ütlus, viimane piisk karikast ja mindi üle piiri.

kõik märgendid liidetult:

<KOK/GÜ/2008/OTP/106> <N><TAH> On olemas </TAH></N> <X><VA> ütlus </VA> </X><I><VMARK> , </VMARK></I> <N><FR><O><OK> viimane piisk <X><VK> karikast </VK> </X></FR> </N> </OK></O> ja <O><OK> mindi üle piiri </OK></O>.

Kui annoteerija ei oska täpset vealiiki määrata, kuid saab aru, millisesse üldkategooriasse viga kuulub, võib vea liigitada konkreetse üldkategooria lahtrisse *muu*, mille märgend kordab sisuliselt veakategooria tähendust, kuid võimaldab tähistada vea piire. Näiteks täpsemalt määratlemata semantika kategooria viga kuulub lahtrisse *muu* ning märgendatakse tähisega (<N><SEM></SEM></N>).

Kui ka vea üldkategooria määratlemine tundub võimatu, võib lause paigutada üldkategooria tasandil lahtrisse *MUUD* (<U><MUU></MUU></U>).

Veakategooriatesse on grupeeritud ainult emakeeleõppijate kirjandites esinenud väärkeelendid. Veatüüpe, mis on küll tõenäolised, kuid mida analüüsitud tekstides ei esinenud, pole kategooriatesse jagatud ega nende jaoks tähiseid loodud. Vajadusel on võimalik veatüüpe lisada ning uute tähistustega märkida või vealiike omakorda väiksemateks allüksusteks jagada.

Klassifikatsiooni loomisel ei ole eristatud vigu ja eksimusi, interlinguaalseid ega intralinguaalseid vigu. Analüüs ei vaatle tekste terviktekstidena ega käsitle teksti

struktuurilist ülesehitust. Vigade tähistamisel ei ole arvestatud õppijate võimalikke erivajadusi.

Soovitav on tekst enne märgendamist lausestada. Sel juhul on hilisemal veaanalüüsil lihtsam vigu sorteerida. Katselisel vigade märgendamisel jäi eelnev lausestamine tegemata, mistõttu keeruliste lausepiire ületavate märgenduste puhul tuli hiljem palju nö musta materjali käsitsi eemaldada. Lause on enamasti piisav kontekst ka sõna stiililise analüüsi jaoks. Vajadusel on võimalik esitada otsitav lause koos eelneva ja järgneva lause või lõiguga.

Vigade märgendamisel on märgendamise kvaliteet olulisem kui kvantiteet.

Vealiikide juurde toodud näited on valitud õppijate tekstidest.

Vigade jagamine kategooriatesse ning veatüüpidesse on tinglik, loodud spetsiaalselt emakeelse õppija tekstis esinevate vigade märgendamiseks ning lähtub kooligrammatika käsitlusest. Klassifikatsioon on aluseks veaanalüsaatori arendamistööle ning lähtepunktiks mitte-eesti emakeelega õppijate vigade märgendamise süsteemile TÜ eesti keele (võõrkeelena) osakonnas.

Tabel 17. Emakeeleõppija korpuse tekstide veakategooriad, veatüübid ja märgendid

veakategooria	tähis	alltüübid	alltüübi tähis
ortograafia	<O></O>	täheortograafiaviga	<O><OK></OK></O>
		aps	<O><APS></APS></O>
		algustäheortograafia viga	<O><AT></AT></O>
		sõna peab olema kirjutatud kokku	<O><KOKKU></KOKKU></O>
		sõna peab olema kirjutatud lahku	<O><LAHKU></LAHKU></O>
		muu	<O><ORT></ORT></O>
inter-punktsioon	<I></I>	liigne kirjavahemärk	<I><LMARK></LMARK></I>
		puuduolev kirjavahemärk	<I><PMARK/></PMARK></I>
		vale kirjavahemärk	<I><VMARK></VMARK></I>
		muu	<I><MARK></MARK></I>
morfoloogia	<M></M>	käändsõnaviga	<M><KS></KS></M>
		pöördsõnaviga	<M><PS></PS></M>
		asesõnaviga	<M><AS></AS></M>
		muu	<M><MORF></MORF></M>
süntaks	<X></X>	väljajätteline lause	<X><PUUDU></PUUDU></X>
		rektsioon	<X><REK></REK></X>
		lauselühend	<X><LL></LL></X>
		järeltäiendi viga	<X><JT></JT></X>
		ühildumisviga	<X><UH></UH></X>
		vale aeg	<X><VT><VT></X>
		vale arv	<X><VA></VA></X>
		vale tegumood	<X><VTM></VTM></X>
		vale kääne	<X><VK></VK></X>
		sõnajärje viga	<X><SNJRG></SNJRG></X>
		muu	<X><SNX></SNX></X>
semantika	<N></N>	tähendus- või sõnavalikuviga	<N><TAH></TAH></N>
		parafraseerimine	<N><PF></PF></N>
		topeldamine	<N><TP></TP></N>
		fraseologismiviga	<N><FR></FR></N>
		reformuleerimisest tulenev segunemine	<N><RF></RF></N>
		muu	<N><SEM></SEM></N>
stiil	<S></S>	argikeelne sõna või väljend	<S><ARGI></ARGI></S>
		sina-stiil	<S><SINA></SINA></S>
		moesõna kasutamine	<S><MOE></MOE></S>
		nominalisatsioon	<S><NOM></NOM></S>
		kordus	<S><KORD></KORD></S>
		grammatiliste vormide kordamine	<S><GKORD></GKORD></S>
		liigne sõna, fraas lauses	<S><LIIG></LIIG></S>
		hakitud stiil	<S><HAK></HAK></S>
		liiga pikk lause	<S><PIK></PIK></S>
		tuleviku väljendamise viga	<S><TUL></TUL></S>
		muu	<S><STIL></STIL></S>
tekst	<T><T/>	vale asesõna kasutamine	<T><VASES></VASES></T>
		ebaselgelt viitav sõna	<T><?VIIT></?VIIT></T>
		kahetimõistetavus	<T><KAH></KAH></T>
		sidususviga	<T><SID></SID></T>
		sõnastusviga	<T><SÕN></SÕN></T>
		sisuviga	<T><SISUL></SISUL></T>

		muu	<T><TXT></TXT></T>
muud	<U></U>	määratlemata vead	<U><MUU></MUU></U>

6.3.4.3. Veatüübid

Veatüüpide illustreerimiseks kasutatatud näidetes on tähistatud ainult need vead, mis esitletavasse tüüpi kuuluvad. Näitelauses esinevaid, kuid mõnda teise kategooriasse või tüüpi kuuluvaid vigu ei ole tähistatud.

6.3.4.3.1. Õigekiri

Õppijakirjandites esinevatest õigekirjaprobleemidest eristatakse ortograafia kategooria all viit tüüpi vigu.

ortograafia	<O></O>	täheortograafiaviga aps algustäheortograafia viga sõna peab olema kirjutatud kokku sõna peab olema kirjutatud lahku muu	<O><OK></OK></O> <O><APS></APS></O> <O><AT></AT></O> <O><KOKKU></KOKKU></O> <O><LAHKU></LAHKU></O> <O><ORT></ORT></O>

1. Täheortograafiavead

ortograafia	<O></O>	õigekirjaviga	<O><OK></OK></O>

<KOK/GÜ/2007/VIL/27/M> Kas rikas riik peab tähendama alati majanduslikku <O><OK> jõukkust </OK></O>?

<KOK/GÜ/2008/OTP/106>Alaealistele on <O><OK> tõkeks </OK></O> vanus.

<KOK/GÜ/2007/VIL/25> Õnneks pole sport ainult oma ala tippude <O><OK> harastus </OK></O>

<KOK/GÜ/2007/VIL/28/M> Uurimistulemuste põhjal on täheldatud, et puberteedieealiste hulgas on suitsetamise populaarus kahanenud tänu mobiiltelefonide <O><OK> lajaldasele </OK></O> levikule.

<KOK/GÜ/2008/OTP/100> Et ajalehte <O><OK> ültse </OK></O> ostetaks, peab lehe kaanel olema mingi põrutav uudis või midagi sellist, mis meeltlahutaks.

eksitud on liidete õigekirjas:

liide *-lik*, *-likkus*

<KOK/GÜ/2007/VIL/5> Soovitatav on mängida korvpalli, mis aitab kaasa ja parandab käelist tegevust, kirjutamist ning annab juurde <O><OK> paindlikust </OK></O>.

<KOK/GÜ/2007/VIL/27/M> Nemad väidavad, et kodumaa põhieesmärgiks peaks saama väga heal tasemel tervishoid, kuid ka selle jaoks on vaja ju <O><OK> majanduslike </OK></O> vahendeid.

liide *-mik*

<KOK/GÜ/2007/VIL/8> <O><OK> Enamiku </OK></O> ajast, mis meile on antud, kasutatakse tavaliselt eesmärkide saavutamiseks ning enese lõbustamiseks.

liide *-ja*

<KOK/GÜ/2007/VIL/21>Milline ideaalne <O><OK> tarbja </OK></O> Lolita oli.

õigekirjavead liite ja võõrsõnatüve liitmisel

<KOK/GÜ/2007/VIL/25> <O><OK> Fännlus </OK></O> on teinud oma töö ja tänapäeval ei ole võimalik tippspordiga tegeleda ilma sponsoriteta.

õigekirjavead liitsõna piiril

<KOK/GÜ/2007/VIL/7> Ometigi on need alad ju pigem <O><OK> vastasoo </OK> pärusmaa.

<KOK/GÜ/2007/VIL/25> Tänu lõõgastavale <O><OK> keskonnale </OK></O> on sageli ka kergem kompromisse leida.

õigekirjavead võõrsõnades:

<KOK/GÜ/2007/VIL/26> Õnneks suudeti ära hoida see <O><OK> insident </OK></O>.

<KOK/GÜ/2008/OTP/116> Eeskujuks valitakse tihti kuulus või andekas näitleja, sportlane või <O><OK> lihdsalt </OK></O> <O><OK> effektne </OK></O>, tuntud ja rikas isiksus.

<KOK/GÜ/2007/VIL/12>Kõik majandusharud on seotud <O><OK> fosiilsete </OK></O> energiaallikatega.

Ortograafiavigade leidmine ning märgendamine on võimalik pool-automaatselt – õigekirjakorrektori abil. Õigekirjakorrektori kasutamisel tuleb arvestada, et speller peab sobivaks ka vormihomonüümia tõttu õigena tunduvaid sõnavorme, seega peab ortograafiavigade parandaja teksti kindlasti ka üle lugema.

vormihomonüümiavead, mida õigekirjakorrektor ära ei tunne

<KOK/GÜ/2007/VIL/25> Iga riik seisab enda <O><OK> kodanikke </OK></O> heaolu eest

<KOK/GÜ/2008/VIL/52> Mõned <O><OK> aastat </OK></O> tagasi olin ma ise üks inimestest, kes kannatasid teatud sorti vägivalla käes.

Tsitaatsõna, tsiteeringu õigekiri

Tsitaatsõnade õigekirja hindamisel arvestatakse nii sõnakuju kui ka esiletõste olemasolu. Tsitaatsõnad, mida ei ole kaldkirjas esile tõstetud, ei vasta ortograafiareeglitele ning need loetakse õpilastekstides veaks.

<KOK/GÜ/2007/VIL/5> Et ratastoolis olev inimene oleks palju julgem, siis tegeletakse Eestis ka neile mõeldud <O><OK> line-tantsuga </OK></O>

<KOK/ PK/2007/VIL/14> <O><OK> Carpe Diem </OK></O>

sõna või lauseosa esiletõstmisel selle tähistamata jätmine

<KOK/GÜ/2007/VIL/26> Hakates defineerima mõistet <O><OK> õnn </OK></O>, siis võin leida sellele vasteid miljardeid

<KOK/GÜ/2007/VIL/27/M> Järelikult ei saa Eesti kohta öelda <O><OK> vaene riik </OK></O>, kuna kõigeks vajalikuks, mille eest riik peaks hoolt kandma, jätkub raha ning jääb veel ülegi.

Speller esitab ka kõik tehnilised ehk teksti sisestamisest tulenevad vead: täht puudu (mehaan*line), tähed vahetatud (**uksus* pro *uskus*, **sutta* pro *surra,** *kiusasus* pro *kiusatus*) vm. Mehaaniliste vigade kohta on tehtud tähelepanek (James 1998: 132), et tähed kipuvad vahetusse minema pigem lühisestes sõnades ning sõna alguses, kuid ära jääma pikemate sõnade teises pooles.

Trükkida soovitav täht läheb tavaliselt vahetusse kas klaviatuuril horisontaalselt, vertikaalselt või diagonaalis asetsevate tähtedega. Näiteks õppijakeele lauses

<KOK/GÜ/2007/VIL/49> Vaatamata sellele usaldatakse <O><APS> saani </APS></O> inimesi pimesid edasi .

on tegemist tüüpilise klaviatuuril diagonaalis asetsevate tähtede vahetamise juhtumiga.

Mõningal juhul on üsna keeruline otsustada, kas õppija ei osanud sõna õigekirja või eksis teksti arvutisse sisestamisel. Kui parandajal jääb kahtlus, kumma veaga tegu on , tuleb viga liigitada pigem ortograafiavigade hulka.

ortograafia	<O></O>	aps	<O><APS></APS></O>

<KOK/GÜ/2007/VIL/32>Täpselt nagu on kindlustatud ühe kõmulise seltskonnatšiki tulevik kuulsa <O><APS> hotteliketi </APS></O> pärijannana.

<KOK/GÜ/2008/OTP/113> Nii see igakuine <O><APS> sissetuleks </APS></O> ära broneeritaksegi ning mida aeg edasi seda suuremaks kasvab jõuetuse tunne arvukate laenukohustuste all.

Algustäheortograafia

ortograafia	<O></O>	algustäheortograafia viga	<O><AT></AT></O>

Suure ja väikese algustähe vead pärisnimedes

<KOK/GÜ/2007/VIL/03> Narkissos, <O><AT> kreeka </AT></O> müütidest pärit kena noormees, murrab nii naiste kui ka meeste südameid

<KOK/GÜ/2008/TRT/86/N/17> Olen õnnelik, et olen saanud elada suurema osa oma elust iseseisvas <O><AT> eesti vabariigis </AT></O>. <O><AT> euroopas </AT></O>

<KOK/GÜ/2007/VIL/29/N> Pärast 2004 aasta maid, mil Eesti ühines Euroopa Liiduga on inimestele teed <O><AT> euroopasse</AT></O> valla.

<KOK/GÜ/2007/VIL/25> Pilte maailmast, kus sõbralikust duellist võib välja kasvada riikidevaheline vaen, näeme läbi televisiooni koos <O><AT> pepsi </AT></O> reklaamiga.

Isikunimede õigekirjavead

<KOK/GÜ/2007/VIL/27/M> Ei ole ilma preemiata jäänud Andrus Veerpalu ega <O><AT> Erkki Nool </AT></O>.

suure ja väikese algustähe vead nimetustes

<KOK/GÜ/2008/OTP/113> Tõsi ta on, et Sakala <O><AT> Keskusest </AT></O> on saamas Uus Sakala ja vanadest lagunevates puumajadest kerkinud <O><AT> Kaksiktornid </AT></O>.

<KOK/GÜ/2007/VIL/25> Raha ringleb, tuju on hea ja vahepeal tõusevad emotsioonid lakke, kuna kaalul on siiski võit teise rahvuse üle - <O><AT> Tüüpiline Maailma Meistrivõistlus </AT></O>.

<KOK/GÜ/2007/VIL/25> Tuleb tõdeda, et <O><AT> Tippsport </AT></O> on muutunud äriks.

osalause alguses on kasutatud suurt algustähte:

> <KOK/GÜ/2007/VIL/15/M> Sama põhimõtet saab aga kasutada väga kavalalt reklaaminduses: <O><AT> Kui </AT></O> sa teatud toodet ei tarbi, siis sa ei kuulu sellesse gruppi.
>
> <KOK/GÜ/2008/VIL/52> Selles koolis oli välja kujunenud omamoodi hierarhia – <O><AT> Vanema </AT></O> klassi õpilased kasutasid nii füüsilist, kui ka vaimset vägivalda nooremate peal
>
> <KOK/GÜ/2007/VIL/25> Ehk siis: <O><AT> Tippu jõudmiseks </AT></O> peab olema mingi eeldus.

Suure algustähe sattumine lause keskel asuva punkti järele võib olla tingitud tagasihoidlikust arvutikasutamisoskusest (programm tõlgendab iga punkti lause lõpetajana ning muudab punkti järel sõna algustähe automaatselt suureks).

> <KOK/GÜ/2008/OTP/97> Umbes 16.-18. <O><AT> Saj </AT></O> . <O><AT> Ei </AT></O> olnud isegi trükimasinaid, millega ajakirju ja lehti trükkida.
>
> <UU/GÜ/2007/VIL/18> ...et tema süüd kuulutada ning eelkõige välja selgitada tõde - st. <O><AT> Rahva </AT></O> puhas uudishimu.

Väikse algustähe kasutamine lause alguses on samuti seletatav väheste kogemustega arvuti kasutamisel.

> <KOK/GÜ/2007/VIL/32> Õpilased omandavad akadeemilisi teadmisi ja ametioskusi, kuid neile ei õpetata rahaga ümberkäimist. <O><AT> võib </AT></O> küll lõpetada kooli väga heade hinnetega, lausa kuldmedaliga, aga finantsküsimustes siiski vaese inimese mõtteviisiga.

Kokku- ja lahkukirjutamine

ortograafia	<O></O>	sõna peab olema kirjutatud lahku	<O><LAHKU></LAHKU></O>

Oma normaalolekus on sõnad üksteisest lahku kirjutatud (Erelt 1997: 76). Kui keelekasutaja kõhkleb, kas kirjutada sõnad kokku või lahku, siis on soovitatud kirjutada need pigem lahku (Erelt 1997: 78).

Kokkukirjutamisvead:

Määrsõnade kokku- ja lahkukirjutamine

kas või kokkukirjutamine

Seda viga on palju ka ajakirjandustekstides, mistõttu on isegi üllatav, et analüüsitud tekstides oli *kas või* kokku kirjutatud vaid kolmel korral, kõikides õppijakorpuse tekstides kokku 43 korral.

> <KOK/GÜ/2007/VIL/6> Igal teisel inimesel kerkib kulm <O><LAHKU> kasvõi </LAHKU></O> veidigi, kui oma tütart või poega lasteaeda viies selgub, et kasvatajaks on mees.

võib-olla kokkukirjutamine

võib-olla ekslik kokkukirjutamine on levinud ka ajakirjandustekstides ning ajakirjanduskorpusest võib leida kümneid sellekohaseid näiteid. (3 viga märgendatud tekstides, 175 viga kõikides praegu õppijakorpuses olevates tekstides.)

> <KOK/GÜ/2007/VIL/14> Järgmistel valimistel lubab seda <O><LAHKU> võibolla </LAHKU></O> keegi teine.

Määrsõnadest kirjutatakse ekslikult kokku mitmed ajaväljendid ning rõhusõnad.

> <KOK/GÜ/2008/OTP/100> <O><LAHKU> Igapäev </LAHKU></O> on lehtedes infot ja meelelahutust. <O><LAHKU> mõnipäev </LAHKU></O>
>
> <KOK/GÜ/2008/VIL/73/N> Televiisorit näeb <O><LAHKU> koguaeg </LAHKU></O> reklaame, mis väidavat, et võta laen ja pea rahulikke jõule.
>
> <KOK/GÜ/2007/VIL/21> Mingil moel mõjutab meedia meid kõiki, <O><LAHKU> kuipalju </LAHKU></O> just –see oleneb inimesest.
> >
> <KOK/GÜ/2007/VIL/5> Et nende kehad oleksid vormis ning <O><LAHKU> normipiirides </LAHKU></O>

kaassõna kokkukirjutamise vead

> <KOK/GÜ/2008/OTP/96> Muutused <O><LAHKU> ajajooksul </LAHKU></O> on loomulikud.
>
> <KOK/GÜ/2008/OTP/96> <O><LAHKU> Minujaoks </LAHKU></O> on ajakirjanduses piisavalt infot ja meelelahutust.

ühend- ja väljendtegusõna õigekirjavead

> <KOK/GÜ/2007/VIL/17> Võib-olla tõesti kui inimesed suudaksid end <O><LAHKU> kokkuvõtta </LAHKU></O>, ning riigid pööraks suuremat tähelepanu keskkonna säästmisele

<KOK/GÜ/2007/VIL/36> Selline stereotüüp on <O><LAHKU> väljakujunenud </LAHKU></O> pikkade aastate jooksul,

mitmusliku vormi kokkukirjutamine järgneva nimisõnaga

<KOK/GÜ/2007/VIL/11> Pärast sellest üle saamist otsustas südi naine, et tema elu pole veel sugugi läbi ja astus linna <O><LAHKU> pensionärideseltsi </LAHKU></O>, mille eakad liikmeid valisid ta sellel aastal endile juhiks.

Lahkukirjutamisvead

ortograafia	<O></O>	sõna peab olema kirjutatud kokku	<O><KOKKU></KOKKU></O>

ne- ja *line*-liiteliste sõnade õigekirjavead

<KOK/GÜ/2007/VIL/3> Kui nooruk saab <O><KOKKU> kuueteist aastaseks </KOKKU></O>, muutub reaalsuseks Teiresiase kuulutus minevikust: „Narkissos on õnnelik vaid siis, kui ta iseennast tundma ei õpi”.

<KOK/GÜ/2007/VIL/17> <O><KOKKU> Keskkonna teemalisi </KOKKU></O> konverentse riikide esindajatele on korraldatud küll.

<KOK/GÜ/2007/VIL/21> Kui <O><KOKKU> maantee äärsetele </KOKKU></O> reklaamplakatitele oli maalitud: “Külastage meie kingitustepoodi”, siis pidid nad seda kohe kindlasti tüdruku nõudmisel tegema.

<KOK/ GÜ/2007/VIL/23> Juba meie <O><KOKKU> aja arvamisest </KOKKU></O> peale on peetud traditsiooniliseks pere, kus eksisteerivad ema ja isa ning palju lapsi

mine-liiteliste nimisõnatuletiste õigekiri

<KOK/GÜ/2007/VIL/8> <O><KOKKU> Vaeva nägemine </KOKKU></O> võib olla sama mõnus kui mõistatused.

<KOK/GÜ/2007/VIL/17> Ilmselt sellisel juhul venitaksime me ainult ökokrahhi <O><KOKKU> pärale jõudmist </KOKKU></O>.

kirjakeele normi alusel kokkukirjutatavad sõnad

<KOK/GÜ/2008/OTP/106> <O><KOKKU> Tihti peale </KOKKU></O> rikkad ja kuulsused arvavad, et neile ei kehti mitmed <O><KOKKU> kirja pandud </KOKKU></O> piirid.

<KOK/GÜ/2008/OTP/106> Öeldakse et siga on must loom, sest ta sööb <O><KOKKU> oma enda </KOKKU></O> väljaheidet.

<KOK/GÜ/2008/VIL/51> Saatus on <O><KOKKU> ala tasa </KOKKU></O> väga kaval ja ettearvamatu.

muud lahkukirjutamisvead

<KOK/GÜ/2007/VIL/23> Teatakse, et isa eemalolek on vajalik pere <O><KOKKU> toime tulekuks </KOKKU></O>.

<KOK/GÜ/2008/OTP/106> Mulle on mulje jäänud, et ka politseinikud mõtlevad, et kuna nemad on seaduse <O><KOKKU> korras hoidjad </KOKKU></O>, võivad nemad mõnda neist eirata.

<KOK/GÜ/2007/VIL/7> Ei ole ainult naised need, kes teevad <O><KOKKU> nii nimetatud </KOKKU></O> meestetöid.

<KOK/GÜ/2007/VIL/29/N> Tänu tööjõu väljaminekule on Eestis paljudes firmades suur <O><KOKKU> tööjõu puudus </KOKKU></O> ning olemasolevad töötajad pole rahul oma töötingimuste ning palgaga.

<KOK/GÜ/2007/VIL/19> Nii umbes kord aastas on küll mõistlik vahetada <O><KOKKU> mobiil telefon </KOKKU></O> ja mõne aasta tagant ka auto uuema vastu, sest vananenud tehnikat on pikema aja möödudes raskem maha müüa,

Liitsõna osa tähistava sidekriipsu puudumine tähistatakse kokkukirjutamise veana.

<KOK/GÜ/2007/VIL/7> Samuti kui oli <O><KOKKU> kartuli panek </KOKKU></O> või <O><KOKKU> võtt </KOKKU></O> ning suur heinategemise aeg olid tervel talurahval käed tööd täis.

Muu viga

ortograafia	<O></O>	muu viga	<O><ORT></ORT></O>

Ajalehtede pealkirjade kirjutamisel topeltmarkeeringu kasutamine

Kui ajakirja nimi on kirjutatud suurtäheliselt, kuid jutumärkides, siis õpilaste kirjandites sellist eksimus veaks ei peeta. Emakeeleõppijakorpuses on markeeringu topeldamine siiski veana tähistatud, kuna markeeringute järgi on hiljem võimalik hinnata, kui hästi on õppijad reegli omandanud ning kas kooligrammatika peab konkreetse teema käsitlemisele suuremat tähelepanu pöörama või ei ole see vajalik.

<KOK/GÜ/2008/OTP/96> Uskuda saab ka <O><ORT><O><ORT> ``Äripäev``-va </ORT></ORT></O>. <O><ORT> ``Eesti Ekspress``</ORT></O> aga jaguneb kaheks.

raamatupealkirja kirjutamisel on järgitud ajakirjamalli

<KOK/GÜ/2007/VIL/41> A.H. Tammsaare teoses <O><ORT> „Tõde ja Õigus” </ORT></O> oli selgelt näha, mida meeleheitel inimene teha võib.

Lühendid

Õpilaskirjandites kui ilukirjanduslikes tekstides tuleks võimalusel lühendite kasutamist vältida. Lühendivigade alla liigituvad nii tarbetud kui ka valesti kirjavahemärgistatud lühendid.

Tarbetud lühendid õpilastekstis

<KOK/GÜ/2007/VIL/25> Antiik-Kreeka <O><ORT> OM </ORT></O> ajal tehti sõdades vaherahu.

6.3.4.3.2. Interpunktsioon ehk vahemärgistus

interpunktsioon	<I></I>	liigne kirjavahemärk puuduolev kirjavahemärk vale kirjavahemärk muu	<I><LMARK></LMARK></I> <I><PMARK/></I> <I><VMARK></VMARK></I> <I><MARK></MARK></I>

Liigne kirjavahemärk

interpunktsioon	<I></I>	liigne kirjavahemärk	<I><LMARK></LMARK></I>

lihtlauses

<KOK/GÜ/2007/VIL/3> Miljonite aastate jooksul evolutsiooni käigus suurenenud ajumaht <I><LMARK> , </LMARK></I> on inimlasest teinud inimese.

koma rindlauses

<KOK/GÜ/2007/VIL/35> Sellistele meestele on olnud eeskujuks nende isad<I><LMARK> , </LMARK></I> ja isadele omakorda nende isad.

lisandi kirjavahemärgistamine

<KOK/GÜ/2007/VIL/31> Algabki duell, millega on edasi antud tõe ja õiguse eest seisva peategelase, prints Hamleti <I><LMARK> , </LMARK></I> võitlus ja hukkumine.

koma võrdluslauses

<KOK/GÜ/2007/VIL/8> Kui piisavalt pinget pakkuval tegevusel ka tulemus on, tõstab see tuju palju enam <I><LMARK> , </LMARK></I> kui telesaade, millest loobusin.

<KOK/GÜ/2008/OTP/100> Ajakirjad on kallimad <I><LMARK> , </LMARK></I> kui ajalehed aga see eest on ajakirjades rohkem lehekülgi ja meelelahutust.

<KOK/GÜ/2008/OTP/116>See on <I><LMARK> , </LMARK></I> kui silmapiiril olev sihtmärk, kuhu tahame jõuda.

liigne punkt lühendi lõpus

<KOK/GÜ/2007/VIL/30> Juba väikesest peale õpib laps selgeks oma nn<I><LMARK> . </LMARK></I> rolli siin elus.

lauselühend lause alguses

<KOK/GÜ/2007/VIL/14> Kõike seda tehes<I><LMARK> , </LMARK></I> peaksin saama pensionipõlves vabamalt ja võimalusterohkemalt elada.

<KOK/GÜ/2007/VIL/26> Minu arvates<I><LMARK> , </LMARK></I> ei tule enamik keskmise palgaga töötajatest vabalt rahaga kuus ots-otsaga kokku.

koma kinnistunud väljendites

<KOK/GÜ/2007/VIL/7> Tugevam sugupool aga eksib arvates <I><LMARK> , </LMARK></I> et naine sellega hakkama ei saa.

<KOK/GÜ/2007/VIL/9> Tema jaoks on see taevane õnn, et ta saab näha oma armastatut õnnelikuna, olgugi <I><LMARK> , </LMARK></I> et temata.

nagu komastamine

<KOK/GÜ/2007/VIL/23> Tänapäeva ühiskonnas on lapsed juba noores eas nutikad ja mõistavad olulisi probleeme<I><LMARK> , </LMARK></I> nagu rahapuudus.

<KOK/GÜ/2007/VIL/9> See on üks tundmustest <I><LMARK> , </LMARK></I> nagu ka armastus - sellel tundel pole kindlat tähendust ega ka põhjust, mis selle tekitada võiks.

kooloni väärkasutus

<KOK/GÜ/2007/VIL/19> Lisaks sellele kasutab inimene veel muidki asju, näiteks <I><LMARK> : </LMARK></I> riideid, ehteid, autot ja arvutit.

<KOK/GÜ/2007/VIL/25> Ettevõtjad otsivad üles oma ala tippvõistlejad ja panevad nood oma toodet promoma, andes vastu võimaluse <I><LMARK> : </LMARK></I> edendada karjääri paremal juhul rikastuda.

Kirjavahemärk puudub

interpunktsioon	<I></I>	puuduolev kirjavahemärk	<I><PMARK>< /PMARK ></I>

põimlause kirjavahemärk

<KOK/GÜ/2007/VIL/1> Seda enam, et kõik läheb järjest kallimaks <I><PMARK> </PMARK> </I> võiksime mõelda, mis on olulisem.

<KOK/ G/2006/TRT/88/M/17>Üks asi, mis teosest välja tuleb <I><PMARK> </PMARK> </I> ja mida ma olen kohanud ka oma elus <I><PMARK> </PMARK> </I> on reeturlikus.

<KOK/GÜ/2007/VIL/6> Kõik sookvoodid ja seadused, mis rõhutavad, et inimesed on võrdsed ja keegi ei ole teistest halvem <I><PMARK> </PMARK> </I> on ju väga toredad, kuid ei maksaks siiski unustada inimese individuaalseid võimeid.

<KOK/GÜ/2007/VIL/9> Kui ma ei elaks seda just nii, nagu ma ise õigeks pean <I><PMARK> </PMARK> </I> ja tahaks kõike lihtsalt kätte saada, siis jätaks õnn mind lõpuks lihtsalt maha.

<KOK/GÜ/2007/VIL/11> Nelja aasta eest, pisut pärast seda <I><PMARK> </PMARK> </I> kui mees oli otsustanud ta maha jätta, sündis tal poeg.

<KOK/GÜ/2007/VIL/2> On noori, kes tahavad elus kuhugi jõuda <I><PMARK> </PMARK> </I> ja noori, kes ei tea või ei hakkagi aru saama, mis on elus tähtsam.

<KOK/GÜ/2007/VIL/5> Kõige silmapaistvamad kuulsused, kes tegelevad spordiga <I><PMARK> </PMARK> </I> on Evelin Ilves ning Arnold Rüütel.

<KOK/GÜ/2007/VIL/6> Sellele oskab vastata vaid iga naine ise, kuid väita, et mehed justkui sunnivad naisi valima <I><PMARK> </PMARK> </I> oleks vale.

<KOK/GÜ/2007/VIL/7> Tugevam sugupool aga eksib <I><PMARK> </PMARK> </I> arvates, et naine sellega hakkama ei saa.

<KOK/GÜ/2008/OTP/101> Võitma õppida on lihtne, aga õppida kaotama <I><PMARK> </PMARK> </I> on ääretult raske.

<KOK/GÜ/2008/OTP/96> Siis on jälle <I><PMARK> </PMARK> </I> mida tuttavatega arutada.

punkt järgarvu järel

<KOK/GÜ/2007/VIL/28/M> HI-viirus hakkas laialdaselt levima 2001 <I><PMARK> </PMARK> </I> aastal, kui viirus levis süstivate narkomaanide hulka.

mõttekriips

<KOK/GÜ/2007/VIL/7> Näiteks lehmi lüpsid ju mõlemad <I><PMARK> </PMARK></I> nii naised kui mehed.

jutumärgid

<KOK/GÜ/2007/VIL/21> Televisioonis on saade nimega <I><PMARK> </PMARK></I> Foorum <I><PMARK> </PMARK></I> kus poliitikud arutavad erinevaid teemasid ning on ka juhtunud, et on hakatud teineteist halvustama.

<KOK/GÜ/2008/OTP/106> On olemas ütlus , <I><PMARK> </PMARK></I> viimane piisk karikast <I><PMARK> </PMARK></I> ja <I><PMARK> </PMARK></I> mindi üle piiri <I><PMARK> </PMARK></I>.

Vale kirjavahemärgi kasutamine

interpunktsioon	<I></I>	vale kirjavahemärk	<I><VMARK></VMARK></I>

otsekõne kirjavahemärgistamine

<KOK/GÜ/2008/OTP/106> On olemas ütlus <I><VMARK> , </VMARK></I> viimane piisk karikast ja mindi üle piiri.

<KOK/GÜ/2007/VIL/7> Paljud minu meessoost sõbrad on mu välja naernud. “Sul ei õnnestu kunagi ühtegi pätt kinni nabida <I><VMARK> . </VMARK></I> ” <I><LMARK> - </LMARK></I> ütlevad nad.

koma mõttekriipsu asemel

<KOK/GÜ/2007/VIL/35> Praegu on selgesti märgatav, et võrdõiguslikkust pole <I><VMARK> , </VMARK></I> see on tegelik ja suur probleem meie arenenud ühiskonnale.

lauselõpumärk

<KOK/GÜ/2007/VIL/36> Kas peale lahutust peaksid lapsed alati emale minema ja mehed peaksid alimente maksma<I><VMARK> . </VMARK></I>

6.3.4.3.3. Morfoloogiavead

morfoloogia	<M></M>	käändsõnaviga pöördsõnaviga asesõnaviga muu	<M><KS></KS></M> <M><PS></PS></M> <M><AS></AS></M> <M><MORF></MORF></M>

Vead on jagatud gruppidesse sõnaliigi kaupa. Õppijatekstide korpuses on praegu eristatud käändsõna vormivead <M><KS>roheliseim</KS></M>, tegusõna vormivead <M><PS>jookstakse</PS></M> ning asesõnade vormivead <M><MORF>millegist</MORF></M>.

Käändsõna muutmisel tekkinud vormivead

morfoloogia	<M></M>	käändsõnaviga	<M><KS></KS></M>

osastava käände moodustamisel tekkinud vead

<KOK/GÜ/2007/VIL/11> Käsitlesin just <M><KS> kuute </KS></M> planeet Maa elanikku 6,5 biljonist.

<KOK/GÜ/2008/VIL/51> Tean ja tunnen <M><KS> paljusi </KS></M> inimesi,kes arvavad,et saavad seda muuta.

<KOK/GÜ/2008/OTP/97> Kui ajakirjanikud ja ajakirjandus niimoodi edasi jätkab ja <M><KS> mõtetuid </KS></M> asju edasi hakkab ajakirjastama, siis muutubki see ajakirjandusvarsti debiilsuseks ja meelelahutuseks

võrdlusastme moodustamisel tekkinud viga

<KOK/GÜ/2007/VIL/39> Ükski paik, isegi siniseim meri ega <M><KS> roheliseim </KS></M> rohi, ei suuda varjutada inimest ning tema panust maailma.

tegusõna muutmisel tekkinud vormivead

morfoloogia	<M></M>	pöördsõnaviga	<M><PS></PS></M>

<KOK/GÜ/2007/VIL/23> Seniks kustutatakse tuli oma väikese käega, <M><PS> jookstakse </PS></M> üle toa voodi poole ning lesitakse voodis rahulolevamalt, kuna materjaalse heaolu pärast ei ole vaja enam mõtteid mõlgutada ning oodatakse põnevaid kingitusi.

<KOK/GÜ/2008/VIL/51> Samas, kui kõik oleks <M><PS> jäätud </PS></M> nii, nagu oli, oleks kõik hästi olnud.

<KOK/GÜ/2007/VIL/23> Siinkohal lepitakse sellega, et üksteist <M><PS> nähtakse </PS></M> harvem, kuid samas suudetakse võimaldada paremat elu.

asesõnaviga

morfoloogia	<M></M>	asesõnaviga	<M><AS></AS></M>

<KOK/GÜ/2008/OTP/97> Tänapäeva ajakirjandus on läinud minumeelest, ja ka paljude teiste inimeste meelest, suisa käest ära, või siis pole ajakirjanikel muust <M><MORF> millegist </MORF></M> kirjutada, kui mõtetutest tegevustest ja teiste inimeste eraelust, mis niikuiniigi kedagi ei huvita.

6.3.4.3.4. Süntaks

süntaks	<X></X>	väljajätteline lause	<X><PUUDU></PUUDU></X>
		rektsiooniviga	<X><REK></REK></X>
		lauselühendi kasutamise viga	<X><LL></LL></X>
		järeltäiendi viga	<X><JT></JT></X>
		ühildumatus	<X><UH></UH></X>
		vale aeg	<X><VT></VT></X>
		vale arv	<X><VA></VA></X>
		vale tegumood	<X><VTM></VTM></X>
		vale kääne	<X><VK></VK></X>
		sõnajärje viga	<X><SNJRG></SNJRG></X>
		sidendi valimise viga	<X><SIDE></SIDE></X>
		muu	<X><SNX></SNX></X>

Väljajättelised laused

süntaks	<X></X>	väljajätteline	<X><PUUDU></PUUDU></X>

Lauseliikmeid võib tekstist välja jätta, kui kontekst annab interpretatsiooniks piisavalt informatsiooni. Kui mõni sõna või lauseosa on lausest välja jäetud, mistõttu lause mõte muutub ebaselgeks või arusaamatuks, on tegemist veaga.

öeldis puudub

<KOK/GÜ/2007/VIL/19> See tähendab, et auto <X><PUUDU> </PUUDU></X> mõnele inimesele vajalik, et tema tööd oleks võimalik teha.

<KOK/ G/2007/VIL/24> Perekonnad, mille isa <X><PUUDU> </PUUDU></X> tööl välismaal, ei ole suuremalt jaolt õnnelikud.

alus puudub

<KOK/ GÜ/2008/OTP/116> <X><PUUDU> </PUUDU></X> Annab võimaluse mingisugust omadust või oskust parandada, et saavutada veelgi paremaid tulemusi.

<KOK/GÜ/2007/VIL/10> Ehk oleks asja veel paremaks teinud <X><PUUDU></PUUDU></X> "ainult" sõnade "ole" ja "rahas" vahel.

täiend puudub

<KOK/GÜ/2008/TRT/150> Riikidel on väga suur roll, et seesuguse ettevõtmise eesmärk jõuaks iga inimeseni. <X><PUUDU> </PUUDU></X> Tähtsuse selgitamiseks reklaamitakse sorteerimist meedia abil.

<KOK/GÜ/2007/VIL/15/M> <X><PUUDU></PUUDU></X> Tulemusena on valdav enamus selle karastusjoogi tarbijatest just noored.

viitav sõna puudub

<KOK/GÜ/2007/VIL/4> Viimases ei maksa aga tingimata süüdistada võrdõiguslikkuse puudumist, vaid pigem võiks seletada <X><PUUDU> </PUUDU></X> asjaolu naissoo väiksema huviga riigivõimu ning poliitika vastu.

<KOK/GÜ/2007/VIL/5> Kelle jaoks mida sport tähendab, sõltub suurel määral <X><PUUDU> </PUUDU></X>, mille nimel seda tehakse

Kõik oleneb <X><PUUDU> </PUUDU></X>, mida õnn iga elusolendi jaoks tähendab.

mittetäielik sidend

<KOK/GÜ/2008/OTP/106> Sisemised piirid võivad sõltuda <X><PUUDU> </PUUDU></X> rassist, kui ka usunditest.

lause on lõpetamata – terve fraas puudub

<KOK/GÜ/2008/OTP/116> Kindlasti ei mõista ma inimesi, kes jäljendavad oma eeskujusid. Käia päevast päeva ringi Elvis Preslina, kanda temale omaseid riideid ja omada täpipealt sarnast soengut <X><PUUDU> </PUUDU></X>.

Rektsiooniviga

Rektsioon ehk sõltumine on lause moodustajate vahel olev grammatiline suhe, mille puhul alistav moodustaja ehk põhi tingib alistuja ehk laiendi grammatilise vormi. Rektsioone võib jagada verbi, substantiivi-, adjektiivi-, adverbi-, kaassõna- ja kvantorirektsiooniks. (Rätsep 1978: 64; EKG II 137–140)

süntaks	<X>	rektsiooniviga	<X><REK></REK></X>

verbirektsioon

<KOK/ GÜ/2008/OTP/101> Pigem puudub oskus kaotada, sest kaotus võrdub <X><REK> häbi </REK></X> ja seda ei tohi ometigi taluda.

<KOK/GÜ/2007/VIL/2> Sageli tahetakse sarnaneda oma <X><REK> iidolile </REK></X>, kes võib kuuluda rokkbändi

<KOK/ GÜ/2007/VIL/5> Kõige kummalisem spordiala on kulturism, mis <X><REK> põhineb ainult jõusaalitreeningutest ning õigest toitumisest </REK></X>.

kaassõnarektsioon

<KOK/GÜ/2007/VIL/7> Miks on aga tänapäeval <X><REK> soolisest võrdõiguslikkusest </REK></X> nii suured ja ägedad arutlused

<KOK/GÜ/2008/VIL/63> Kuid mis saaks siis, kui me näeks ette enda <X><REK> saatuste </REK></X> ja <X><REK> valikud </REK></X> ning seejärel sekkuks enda saatusesse ja teeks vale valiku?

<KOK/GÜ/2008/OTP/101> Kuid selleks, et jõuda võitudeni, on mul väga palju <X><REK> kaotused </REK></X> ja tagasilööke vaja olnud.

Lauselühendi vead

süntaks	<X>	lauselühend	<X><LL></LL></X>

<KOK/GÜ/2007/VIL/35><X> <LL> Lugedes uudiseid liiklusõnnetustest </LL></X> on peaaegu iga õnnetuse põhjustaja mees, kuid millegipärast peetakse naisi alati kehvadeks autojuhtideks.

<KOK/GÜ/2007/VIL/11><X> <LL> Saades reaalsuselt karmi hoobi </LL></X>, üritas ta end kokku võtta, kuid siiani kimbutab viinakurat, kes ei anna talle luba inimlikuks eluks.

<KOK/GÜ/2007/VIL/9><X> <LL> Hästi minnes </LL></X>, moodustab kogu see kooslus inimese rahulolu, tema õnne, mis alateadlikult ongi iga inimese elu eesmärk.

<KOK/GÜ/2008/VIL/63>Kas tulevikku <X><LL> ette nähes </LL></X> võib meid tabada sama saatust, mis tabas kuningas Oidipust?

Järeltäiendi viga

süntaks	<X>	järeltäiend	<X><JT></JT></X>

<KOK/GÜ/2007/VIL/39>Need erinevad üksteisest väga jõuliselt, kuid ühendavad endis <X><JT> olulisuse inimgruppidele </JT></X>

<KOK/GÜ/2007/VIL/3> <X><JT> Põhjus hirmsale needusele </JT></X>, mis ohvrit tabas, oli ärapõlatud armuvaluskannataja kättemaksuiha.

<KOK/GÜ/2007/VIL/5> Isegi väiksematesse kohtadesse ehitatakse juurde erinevaid jõusaale ning ujulaid, et kõikidel oleks <X><JT>kergem ligipääs sporditarvetele </JT></X>.

Ühildumisviga

Ühildumise all mõeldakse teatud vormitunnuste tingitust teise elemendi (sõna või moodustaja) vastavatest vormitunnustest või semantilistest omadustest. Kui põhja grammatiline vorm ja/või grammatiline tähendus tingib laiendi samasuguse grammatilise vormi, siis on tegemist ühildumisega. (EKK: 376)

süntaks	<X></X>	ühildumine	<X><UH></UH></X>

ühildumisviga

<KOK/GÜ/2007/VIL/9> Suutsin hingede nägusid lugeda kui <X><UH> avatuid </UH></X> raamatuid, mis sisaldasid nende elulugusid, seega alustasin sisude tutvumisega .

<KOK/GÜ/2008/VIL/51>Nagu oli näha Sophoklese trägöödias <X><UH> "Kuningas Oidipuses" </UH></X>, ei suudetud saatust muuta.

Vale ajavormi kasutamine

süntaks	<X></X>	vale ajavorm	<X><VT></VT></X>

<KOK/GÜ/2007/VIL/11> Ta on rõõmus, kui saab iga päev palli taguda ja see teeb tasa kogu vaeva, mis ala avastamise jaoks vaja <X><VT> läks </VT></X>.

<KOK/GÜ/2007/VIL/65> Kui on ette kirjutatud halb saatus ja sul <X><VT> õnnestus </VT></X> seda muuta, on sinu saatus veel hullem.

<KOK/GÜ/2007/VIL/35> Naised ja mehed ei pea muutuma ühesugusteks, kuid nende õigused, kohustused ja võimalused ei tohi sõltuda sellest, kas nad <X><VT> sünnivad </VT></X> naise või mehena.

Vale arvu kasutamine

süntaks	<X></X>	vale arv	<X><VA></VA></X>

<KOK/GÜ/2008/OTP/106> Muusikas peetakse <X><VA> piirideks </VA></X> häält.

<KOK/GÜ/2008/OTP/106> On olemas <X><VA> ütlus </VA></X> , viimane piisk karikast ja mindi üle piiri.

Vale tegumoe kasutamine

süntaks	<X></X>	vale tegumood	<X><VTM></VTM></X>

<KOK/GÜ/2007/VIL/5> Varasematel perioodidel <X><VTM>pühendati</VTM></X> tervislikule eluviisile rohkem aega oma kiirest elutempost, kui praegu.

Käände valimise viga

süntaks	<X></X>	vale kääne	<X><VK></VK></X>

<KOK/GÜ/2007/VIL/19> Majanduslikult erineval astmel olevad inimesed ostavad <X><VK> erinevas hinnaklassis </VK></X> tooteid.

<KOK/GÜ/2005/TRT/89/M/16> Thomas oli tavaline käsitööline, kui ta <X><VK> üks päev </VK></X> teada sai, et tema kauge sugulane on ära surnud.

Sõnajärje viga

süntaks	<X></X>	sõnajärg	<X><SNJRG></SNJRG></X>

Rõhusõnaga tähistatud lauseosa peab paiknema vahetult rõhutatava ees (Kasik 2007: 114) Määratlev, rõhutav, täpsustav lauseosa peab paiknema rõhutatava, määratletava või täpsustatava ees.

<KOK/GÜ/2007/VIL/2> Noorte seas on <X><SNJRG> ka populaarseks saanud </SNJRG></X> alkohol, peod ja narkootikumid ning neid kätte saada on küllaltki lihtne.

<KOK/GÜ/2007/VIL/8> Kuigi tänapäeva noortele võib tunduda, et ilma arvuti ja telekata pole tegevust, said me esivanemad <X><SNJRG> ilma hakkama vägagi edukat </SNJRG></X>

määratlevad (hulga)sõnad ei paikne fraasis esimesel kohal

<KOK/ PK/2007/VIL/14>Võin hetkel küll mõelda, et suurena tööl käies hakkan <X><SNJRG> raha hulgaliselt </SNJRG></X> kokku hoidma, et kunagi oleks hea seda kasutada, kuid vaevalt.

<KOK/GÜ/2007/VIL/6>Isapuhkused on <X><SNJRG> väga teretulnud tänapäeval </SNJRG></X>

muud sõnajärjevead

<KOK/GÜ/2007/VIL/6> <X><SNJRG> Mis teeb probleemist tegeliku või pseudoprobleemi on inimeste vaatenurk </SNJRG></X> , st kas ühiskond leiab, et sellest tuleks numbrit teha või mitte.

<KOK/GÜ/2007/VIL/10> Usun et Eesti riigis jõuaks ajalehekaantele 5 miljoni kroonise Mercedes-Benzi ostmise või mõne rahvusvahelise staari <X><SNJRG> kutsumisega erapeole </SNJRG></X>.

<KOK/GÜ/2007/VIL/3> Arukas Sinuhe õppis oma vigadest ning saavutas <X><SNJRG> kohanedes tavadega </SNJRG></X> edu haridusteel.

<KOK/GÜ/2007/VIL/17> Tekstid nendelt on väga ahvatleva sisuga ja <X><SNJRG> kutsuvad tihti kutsuvad ostma </SNJRG></X>

<KOK/GÜ/2008/OTP/115> <X><SNJRG> Sellest summast mingi protsent antakse ära </SNJRG></X>.

Muu

süntaks	<X></X>	muu	<X><SNX></SNX></X>

Sõnakordust käsitletakse stiili all.

Viitamisvigu käsitletakse teksti all.

Sõnajärjest tulenevaid kahetimõistetavusi käsitletakse semantika all.

6.3.4.3.5. Semantika

Kirjandite semantilised eksimused on peamiselt kahte tüüpi:

1) konkreetset sõna haaravad vead;

2) tervet fraasi haaravad vead.

Semantika	<N></N>	tähendus- või sõnavalikuviga parafraseerimine topeldamine fraseologismiviga reformuleerimisest tulenev segunemine muu	<N><TAH></TAH></N> <N><PF></PF></N> <N><TP></TP></N> <N><FR></FR></N> <N><RF></RF></N> <N><SEM></SEM></N>

1. Tähendusviga või sõnavalikuviga

semantika	<N></N>	tähendusviga	<N><TAH></TAH></N>

paronüümiviga

<KOK/GÜ/2007/VIL/41>Esialgu on siiski vaja <N><TAH> teavitada ennast </TAH></N>

<KOK/GÜ/2007/VIL/41> Kui keegi on sõbralt laenu võtnud ja <N><TAH> tähtpäev </TAH></N> läheneb, aga raha ei ole, siis ta võib hakata meeleheitlikult raha otsima.

<KOK/GÜ/2007/VIL/12> <N><TAH> Peale </TAH></N> orkaan Katariina hävitus tööd ja sõda Iraagis tõusid kütuste hinnad hüppeliselt.

valesti valitud sõna

<KOK/GÜ/2007/VIL/2> Linnalapsed on ka sellised uhked, ülbed ja üldse kõrgete <N><TAH> väärtushinnangutega </TAH></N> <N><TAH> kõrge eneseväärikusega </TAH></N>).

<KOK/GÜ/2006/VIL/48>Elu on näidanud, et varem või hiljem avastatakse aga <N><TAH> enamus </TAH></N> dopingu juhtumeid.

olematu sõna kasutamine

<KOK/GÜ/2008/OTP/96> Igasugu koledad <N><TAH> emodeemonid </TAH></N> päästetakse valla.

hüponüümi valimine

<KOK/GÜ/2008/OTP/113> Mida rohkem suudetakse teenida seda rohkem ka endale lubada ning õige pea avastataksegi end kandvat väsinud Marati dresside asemel kalli <N><TAH> firmamärgi </TAH></N> logoga ehitud riideid, mis on juba enne poeletile sattumist ära kulutatud ja auke täis.

<UU/GÜ/2007/VIL/18> Nii on ka netikommentaarides Estonia teemal - inimeste jaoks on <N><TAH> internetikommentaar </TAH></N> kohaks, kus on võimalik oma arvamus välja öelda

hüperonüümi valimine

<KOK/GÜ/2008/OTP/100>Et ajalehte ültse ostetaks, peab <N><TAH> lehe kaanel </TAH></N> olema mingi põrutav uudis või midagi sellist, mis meeltlahutaks.

kollokatsiooniline viga

<N><SISUL> matuserongi möödudes <N><TAH> langetatakse müts </TAH></N>

Parafraseering

semantika	<N></N>	parafraseering	<N><PF></PF></N>

<KOK/GÜ/2007/VIL/41> Kõige tähtsam on inimese juures sisemus aga, kui see on täis meeleheidet ja pettumust oma välimuse vastu, siis see mõjutab tema olemust. <N><PF> Iga mitterahul omadus </PF></N> teeb tusaseks ja kurvaks, mida rohkem neid koguneb, seda kurvemaks muutub inimene.

<KOK/GÜ/2007/VIL/2> Arvan ise, et linnalapsed oskavad paremini suhelda, kuna puutuvad iga päev kokku <N><PF> palju rohkem erinevate </PF></N> inimestega

<KOK/GÜ/2007/VIL/5> Nägin üht dokumnetaalfilmi, mis rääkis 50-aastastest ning <N><PF> sellest vanusest veel vanematest inimestest </PF></N>, kes tegelesid selle alaga.

<KOK/GÜ/2007/VIL/5>Maailmas eksisteerib palju <N><PF> mitte lihasööjaid </PF></N>

<KOK/GÜ/2007/VIL/23>Olles mitmesaja kilomeetri kaugusel tööl, nähakse pereliikmeid küll mitme nädala tagant, kuid töötatakse vähem ja <N><PF> kodumaal olev aeg </PF></N> pühendatakse kodustega aega viites.

Topeldamine

semantika	<N></N>	topeld	<N><TP></TP></N>

<KOK/GÜ/2007/VIL/37> Inimene on <N><T> arenenud välja </T></N> loomast, kes on alati elanud ühe tähtsa reegli järgi – tugevam jääb ellu.

Fraseologismiviga

Fraseologismi mittetundmisest tingitud viga.

semantika	<N></N>	fraseologismiviga	<N><FR></FR></N>

<KOK/GÜ/2008/OTP/106> On olemas ütlus, <T><FR> viimane piisk karikast </FR></N> ja mindi üle piiri.

Reformuleerimisest tulenev segunemine

semantika	<N></N>	reformuleerimisest tulenev segunemine	<N><RF></RF></N>

Sulam – reformuleerimisest tulenev segunemine, mida on nimetatud ka kontaminatsiooniks (James 1998: 111). Segunemine on omane suulisele kõnele. Eesti õpilaste tekstides ei ole sellist nähtust analüüsitud. Sulamite põhjuseks on peetud kahe semantiliselt seotud struktuuri üheaegset aktiveerumist, mille vahel tuleks valida, kuid kirjutaja/kõneleja ei vali, vaid segab kaks struktuuri omavahel. Sulamitest eristatakse kolme tüüpi:

1) ühendamise tulemusel jääb lausest otsekui sõna puudu, st ei valita kumbagi. 2) valitakse mõlemad, st topeldatakse mingit väljendi osa, 3) võetakse kummastki struktuurist kasutusele üks osa)(James: 1998: 111; pikemalt Nesselhauf 2005).

<KOK/GÜ/2008/OTP/116> <N><RF> Kuid inimeste soov on eeskujule niivõrd väliselt sarnaneda. </RF></N>

soov eeskujuga väliselt sarnaneda + nii palju kui võimalik

<KOK/GÜ/2008/OTP/115> Mõned inimesed <N><RF> peavad päevad läbi rassima rasket tööd, et raha saada </RF></N>.

peavad tegema rasket tööd + peavad tööl rassima

<KOK/GÜ/2007/VIL/22> Masaaki Hatsumi on tuntud võitluskunstide meister, ta on üle 70 aasta vana, kuid <N><RF> suudab vaevalt jõudu kasutamata </RF></N> jagu saada ka palju noorematest vastastest.

suudab jõudu kasutamata + suudab (ilma) vaevata

<KOK/GÜ/2007/VIL/23> Kui mees lahkub kodust, et teenida perele lisaraha, <N><RF> jätab ta naise õlgadele kanda raske koorma. </RF></N>

jätab naise õlgadele koorma + jätab koorma naise kanda

<KOK/GÜ/2007/VIL/2> On noori, kes tahavad elus kuhugi jõuda ja noori, <N><RF> kes ei tea või ei hakkagi aru saama </RF></N>, mis on elus tähtsam.

kes ei tea ega tahagi teada + ei hakka aru saama

<KOK/ PK/2007/VIL/14>Terve elu säästlikult elades, palju tööd rügades ja õigeid valikuid tehes saaksin tulevikus jalad seinale visata ning enda <N><RF>tulemuste vilju </RF></N> nautida.

töö tulemusi + töövilju

<KOK/GÜ/2007/VIL/26> Minu arvates, <N><RF> ei tule enamik keskmise palgaga töötajatest vabalt rahaga kuus ots-otsaga kokku. </RF></N>

ei tule vabalt rahaga välja + ei tule ots-otsaga kokku

6.3.4.3.6. Stiil

Stiil on suhtlussituatsioonile ja -eesmärgile vastav keelekasutusviis. Harilikud, tavapärased standardsed keelendid sobivad oma neutraalsuses igale poole ega tekita kuskil tõrkeid. Stiilivärvinguga sõnad ja väljendid hälbivad tavakeelest, on ebatavalised, rõhutatud ning väljaspool oma stiili tekitavad stiilimõra. (EKK: 508)

stiil	<S></S>	argistiil	<S><ARGI></ARGI></S>
		sina-stiil	<S><SINA></SINA></S>
		moesõna kasutamine	<S><MOE></MOE></S>
		nominalisatsioon	<S><NOM></NOM></S>
		kordus	<S><KORD></KORD></S>
		gramm vormide kordamine	<S><GKORD></GKORD></S>
		liigne sõna, fraas lauses	<S><LIIG></LIIG></S>
		hakitud stiil	<S><HAK></HAK></S>
		liiga pikk lause	<S><PIK></PIK></S>
		tuleviku väljendamise viga	<S><TUL></TUL></S>
		muu	<S><STIL></STIL></S>

Stiilivigade hulka kuuluvad

registrivalikuvead

argikeele kasutamine

a) argikeele sõnad (väljendeid) – keelekuju, mida kasutatakse vähenõudlikumas igapäevases suhtlemises. Kirjakeelega võrreldes on argikeel normilõdvem. (Erelt 1997: 510–511). Argikeelde kuuluvate sõnade ja väljendite veatähis õppijakeelekorpuses on <S><ARGI> pensieas </ARGI></S>.

stiil	<S></S>	argi	<S><ARGI></ARGI></S>

<KOK/GÜ/2007/VIL/13> Ülikoolis tahaksin õppida arhitektuuri, sest mulle meeldib joonestamine ja matemaatika ning üldse <S><ARGI> kõiksugu </ARGI></S> asjade konstrueerimine ja ehitamine.

<KOK/GÜ/2007/VIL/19> Vanad ja veel küllaltki <S><ARGI> kobedad </ARGI></S> asjad jäävad seisma, uued ja veel <S><ARGI> kobedamad </ARGI></S> hõivavad nende koha.

<KOK/GÜ/2007/VIL/9> Mehel <S><ARGI> oli vedanud </ARGI></S>, et ta kasvas sellises peres, kuid lõpuks saab ka <S><ARGI> vedamine </ARGI></S> otsa.

<KOK/GÜ/2008/OTP/96> <S><ARGI> Noh, las olla </ARGI></S>, elu on <S><ARGI> siuke </ARGI></S>.

<KOK/GÜ/2007/VIL/7> Mulle näib, et inimestel on igav, elu on liiga mugavaks läinud ja siis on vaja leida probleeme, millest <S><ARGI> kära teha </ARGI></S>.

b) *sina*-stiili kasutamine viitab samuti argisele kõnepruugile ning kirjandis peetakse seda stiiliveaks. Pronoomenite statistiline analüüs näitas, et gümnaasiumiealised õpilased oskavad kirjandites üsna hästi *sina*-stiili vältida, kuid põhikooliõpilaste tekstides oli asesõna *sina* kvantorite ja pronoomenite koondtabelis sageduselt kümnes.

stiil	<S></S>	*sina*-vormi kasutamine	<S><SINA></SINA></S>

Veamärgendiga tuleb tähistada nii lühike pronoomen *sa,* pikk pronoomenivorm *sina* kui ka kõik selle pronoomeni käändelised vormid.

<KOK/GÜ/2007/VIL/26> <S><SINA> Maailm muutub, sa hakkad nägema enda ümber toimuvat uue vaatenurga alt </SINA></S>.

<KOK/GÜ/2007/VIL/9> Kõige sellega kaasnevad inimesed, kes hakkavad <S><SINA> sind </SINA></S> edaspidideses elus aitama, toetama või takistama.

sina-stiili märgend tuleb lisada ka ainsuse 2. pöördes olevatele verbidele.

<KOK/GÜ/2007/VIL/26> <Sellest hetkest on tunne, nagu <S><SINA> avaksid </SINA></S> oma silmad taas esimest korda.

Kui pronoomen *sa* ning ainsuse 2. pöörde vormis verb asuvad lauses kõrvuti, on otstarbekas lisada neile ühine märgend.

<KOK/GÜ/2008/VIL/52><S><SINA> Tavaliselt, kui sa näed väiksemaid lapsi rusikatega vehklemas, siis on ju loogiline, et sa lähed vahele ja ütled </SINA></S>, et vägivald pole lahendus.

Moesõnade (-väljendite) kasutamine

stiil	<S></S>	moesõna	<S><MOE></MOE></S>

<KOK/GÜ/2007/VIL/12> Kui biodiisel läheb diisliga võrreldes odavamaks hakatakse seda <S><MOE> koheselt </MOE></S> laialdaselt kasutama ja kiitma

<KOK/ GÜ/2007/VIL/8> Õndsad on need, kes on vaimust vaesed, kuid kui sellist omadust antud pole, siis on võimatu ka <S><MOE> läbi </MOE></S> narkootikumide elujärge parandada.

<KOK/GÜ/2007/VIL/13> Ülikoolis tahaksin õppida arhitektuuri, sest mulle meeldib joonestamine ja matemaatika ning üldse kõiksugu <S><MOE> asjade </MOE></S> konstrueerimine ja ehitamine.

<KOK/GÜ/2007/VIL/12> <S><MOE> Hetkel </MOE></S> naudime veel viimaseid nafta kuldaja vilju, odavat ja kiiret transporti,

<KOK/GÜ/2007/VIL/36> Naised, kes <S><MOE> on võrdõiguslikkuses kinni </MOE></S>, võivad arvata seda, et mees näitab oma üleolekut.

<KOK/GÜ/2007/VIL/39> <S><MOE> Igapäevaselt </MOE></S> on uudisteportaalides levinud teemadeks tööstus, arvutid ning tehnika.

<KOK/GÜ/2007/VIL/44> Eestit rõhuti nii saksa kui ka vene võimu poolt ja siis oli <S><MOE> praktiliselt </MOE></S> võimatu vaimuelu arendada.

<KOK/ GÜ/2007/VIL/26> <S><MOE> Tänu </MOE></S> sellisele probleemile, mis on saavutatud, hakkavad inimesed õnne võrdsustama rahaga.

<KOK/GÜ/2007/VIL/32> <S><MOE> Tegelikult </MOE></S> on väga rumal arvata, et tänapäevase koolisüsteemi pakutav haridus valmistab noori ette eluks, mis neid pärast kooli lõpetamist ootab.

kantseliitlike moesõnade kasutamine

<KOK/GÜ/2007/VIL/33> Kuid õppevahendeid ei ole vaja endal <S><MOE> omada </MOE></S>.

<KOK/GÜ/2007/VIL/4> Just seetõttu on ääretult oluline <S><MOE> tagada </MOE></S> kõikidele inimgruppidele võrdsed õigused ning mittediskrimineerimine.

<KOK/GÜ/2007/VIL/41> Ta peaks üle saama meeleheitest, mis tal autokooliga <S><MOE> kaasneb </MOE></S> ja suhtuma kõigesse rahulikumalt.

<KOK/GÜ/2007/VIL/22> Muidugi on mõlemad asjad <S><MOE> inimese poolt </MOE></S> välja mõeldud

<KOK/GÜ/2007/VIL/28/M> Suitsetajate ja piraatlusega <S><MOE> seoses </MOE></S> oleme jõudnud samuti Euroopas esikümnesse.

Nominalisatsioon

Nominalisatsiooniks nimetatakse sündmust väljendavaid nimisõnafraase, mille tuumaks on teonimi, nt *tubade kraamimine*, või muu verbist tuletatud nimisõna, nt *majanduse kiire areng* (Kasik 2007: 136). Õigekeelsuslikult probleemikas nominaalstiil on üsna levinud. Temast on kujunenud erilise võimukeele, nn kantseliidi üks vahendeid: tuletatakse seal, kus tegusõna hõlbustaks teksti jälgimist ja mõistmist (Kerge 2005: 38).

stiil	<S></S>	nominalisatsioon	<S><NOM></NOM></S>

<KOK/GÜ/2007/VIL/4> Enam ei tekita hämmingut ka <S><NOM> isa jäämine lapsehoolduspuhkusele </NOM></S>.

<KOK/GÜ/2007/VIL/03> Ühiskonnas <S><NOM> on informatsiooni omamine määravaks teguriks eelispositsiooni saavutamisel </NOM></S>.

<KOK/GÜ/2007/VIL/12> Sellega aga <S><NOM> kaasneb tarbijate ostuvõime langus </NOM></S> ja kuna kaupasid ei osteta piisavalt peaksid paljud ettevõtted tootmise lõpetama ja vallandama töötajaid

Kordus

Juhuslikult lähestikku asuvad ühesugused sõnad või sõnavormid muudavad lause kohmakaks (Kasik *et al.* 2007: 131). Sõnakordused on analüüsitavad

ainult kontekstis. Juhuslikust kordamisest või sõnavara ahtusest tingitud kordamist tuleb eristada rõhutamiseks kasutatud sõnakordusest.

stiil	<S></S>	kordus	<S><KORD></KORD></S>

liiased sõnakordused

lausestruktuuri kordamine järjestikustes lausetes

<KOK/GÜ/2007/VIL/36> <S><KORD> On selliseid, kes tunnevad </KORD></S>, et nad peavad kõike kontrollima. <S><KORD> Mehed, kes tunnevad </KORD></S> ennast naistega võrdselt. <S><KORD> Mehed, kes lasevad </KORD></S> naistel enda eest kõike kontrollida ja korraldada.

sõna, sõnatüve või väljendi kordamine

<KOK/GÜ/2007/VIL/5> Kehalise kasvatuse tund annab lühikese ülevaate erinevatest spordialadest, mis tuleb kasuks neile, kes soovivad millegagi <S><KORD> tegeleda </KORD></S>, et saaks endale õige <S><KORD> tegevuse </KORD></S> leida.

<KOK/GÜ/2007/VIL/41> Mida rahulikum <S><KORD> ta </KORD></S> on ja mida vähem <S><KORD> ta </KORD></S> unistab eksami sooritamisest, seda suurem on <S><KORD> tal </KORD></S> lootus eksam edukalt sooritada.

sidesõna kordamine

<KOK/GÜ/2007/VIL/11> Muidu ta ei lase oma pisaratel nii vabalt voolata, kuna tirtsul on kodus olemas kass, oma tuba, mis on üleni roosa <S><KORD> ja </KORD></S> äärest ääreni mänguasju täis <S><KORD> ja </KORD></S> mille koristamisega ta kunagi vaeva ei pea nägema, sest isa palkas selleks tööks eraldi inimese.

<KOK/GÜ/2008/OTP/101> Tegeledes spordiga on tulemused minu jaoks vägagi aktuaalsed <S><KORD> ja </KORD></S> loomulikult ma võistlen <S><KORD> ja </KORD></S> treenin võidu nimel.

asesõna kordamine

<KOK/GÜ/2007/VIL/40> <S><KORD> Meie </KORD></S> elu jooksul saadavad <S><KORD> meid </KORD></S> <S><KORD> meie </KORD></S> sõbrad ja <S><KORD> meie </KORD></S> emotsoonid, kes <S><KORD>meile </KORD></S> valikute tegemisel nõu annavad.

<KOK/ PK/2007/VIL/14> Võiksin ju kõike <S><KORD> seda <S><KORD> mõelda, aga siiski ei tee <S><KORD> seda </KORD></S>.

<KOK/ PK/2007/VIL/14> Südames <S><KORD> ma </KORD></S> usun, et vanas eas ei vaja <S><KORD> ma </KORD></S> põnevust ega ka seiklusi.

rõhusõna kordamine

<KOK/GÜ/2008/VIL/53> Inimeste manipuleerimiseks kasutatakse <S><KORD> ka </KORD></S> peale ajalehtede <S><KORD> ka </KORD></S> televisiooni ja <S><KORD> ka </KORD></S> raadiot.

Grammatiliste vormide kordamine

Grammatiliste vormide kordamine muudab väljenduse kohmakaks.

stiil	<S></S>	grammatiliste vormide kordamine	<S><GKORD></GKORD></S>

sarnaste käändevormide kordamine

<KOK/GÜ/2007/VIL/5> Ratastooliinimestele on väga tähtis tegeleda kergema <S><GKORD> raskustasemega võimlemisega </GKORD></S>, mille tulemusel võib uuesti käimisvõimeliseks persooniks saada.

<KOK/GÜ/2007/VIL/11> Kuid mehega isekeskis on nad arutanud, et sooviksid minna kahekesi koos <S><GKORD> nädalasele puhkusereisile Kreetale </GKORD></S> , mis läheks maksma täpselt nii palju, et nad praegu seda endale lubada ei saa.

pronoomeni kasutamine kõrvuti pöördelõpuga

<KOK/GÜ/2007/VIL/17> Ilmselt sellisel juhul <S><GKORD> venitaksime me </GKORD></S> ainult ökokrahhi pärale jõudmist.

<KOK/GÜ/2007/VIL/15/M> <S><GKORD> Elame me </GKORD></S> ju ajastul, mil räägitakse alternatiivsest energiast ja materjalide korduvkasutamisest, kuna maavarad on otsa lõppemas ja prügimäed on täis.

Liigne sõna, fraas lauses

stiil	<S></S>	Liigne sõna, fraas lauses	<S><LIIG></LIIG></S>

rõhusõna

<KOK/GÜ/2007/VIL/2> Paljud linnalapsed on <S><LIIG> ka </LIIG></S> mitmesuguste kompleksidega, mida on eriti noormeestel.

<KOK/ PK/2007/VIL/14> Südames ma usun, et vanas eas ei vaja ma põnevust ega <S><LIIG> ka </LIIG></S> seiklusi.

<KOK/GÜ/2008/OTP/97> Varsti võibolla muutubki ajakirjandus <S><LIIG> juba </LIIG></S> meelelahutuseks

muu liigne

<KOK/ G/2007/VIL/8> Näiteks, kui tõotatud maad otsides reeta sõpru ja riskida kõigega võib saatus olla palju karmim kui ettevõtmise eel. Kui <S><LIIG> seda </LIIG></S> heaolu nimel röövida kellegi teise õnn, siis polegi nii kerge saadud süümepiinadega elada.

<KOK/GÜ/2008/VIL/52> Ja lõpuks ei astunudki õpetajad <S><LIIG> selle </LIIG></S> vahele, vaid hoopis teenindav personal - söögitädid.

kõrvalekaldumised teksti üldisest stiilist.

<KOK/GÜ/2007/VIL/13> <S><LIIG> Heakene küll, niipalju siis sellest </LIIG></S>.

Hakitud stiiliga laused

Hakitud stiiliga lause on sidesõnaga algav kõrvallause, mis on tehtud omaette lauseks või lauselühendiga algav lause, kus puudub pöördeline verbivorm (Habicht 2008).

stiil	<S></S>	hakitud lause	<S><HAK></HAK></S>

<KOK/GÜ/2008/VIL/51> <S><HAK> Samas, kui kõik oleks jäätud nii, nagu oli, oleks kõik hästi olnud. Ja vastupidi. Inimesed aga ei tea seda ette. Ja ehk nii ongi parem. </HAK></S>

Liiga pikk lause

Liiga pikaks võib pidada lauset, mis koosneb keerulise ülesehitusega, omavahel põimunud ning halvasti seotud lausetest. Tavaliselt põhjustavad liiga pikad laused õppijatekstis kirjavahemärgistamise vigu. Gümnaasiumiõpilase keskmine lausepikkus on 13 sõna.

stiil	<S></S>	liiga pikk lause	<S><PIK></PIK></S>

42 sõna pikkune lause

<KOK/GÜ/2007/VIL/35> <X><PIK> Eeldan, et hooldusõiguse saab see vanem, kes on rohkem lapsega koos olnud, kui lahutuse käigus isa avastab, et tahab nüüd isa olla, aga kooselu ajal pidas kõike muud tähtsamaks, siis tuleks kindlasti talle anda võimalus lapsega kohtuda, aga ema on siiski ema. </PIK></S>

49 sõna pikkune lause

<KOK/GÜ/2007/VIL/43> <X><PIK> On ometi eeskujulikke näiteid, kes on ennast suutnud Euroopas juba tõestada: kirjanduses Jaan Kross; kunstis Kalev Mark Kostabi, kes on öelnud, et ei taha olla mitte ainult ühes reas Leonardo da Vinci, Michelangelo või Picassoga, vaid soovib seal olla esimene; draamakunstis on ilma teinud Elmo Nüganen oma lavastustega jne. </PIK></S>

Tuleviku väljendamise viga

süntaks	<S></S>	tulevik	<S><TUL></TUL></S>

<KOK/GÜ/2007/VIL/37> Väärteod peavad <S><TUL> saama karistatud </TUL></S> , kuid karistused ei tohi olla liiga leebed ega ka mitte liialt karmid.

<KOK/ GÜ/2008/OTP/106> Kes selle paragrahvi ületab, <S><TUL> saab karistatud </TUL></S>.

stiil	<S></S>	muu	<S><STIL></STIL></S>

6.3.4.3.7. Tekstivead

Tekst	<T><T/>	vale asesõna kasutamine ebaselgelt viitav sõna kahetimõistetavus sidususviga sõnastusviga sisuviga muu	<T><VASES></VASES></T> <T><?VIIT></?VIIT></T> <T><KAH></KAH></T> <T><SID></SID></T> <T><SÕN></SÕN></T> <T><SISUL></SISUL></T> <T><TXT></TXT></T>

Vale asesõna kasutamine

Õppija viitab entiteedile vale pronoomeniga.

Tekst	<T><T/>	vale asesõna	<T><VASES></VASES></T>

<KOK/ G/2007/VIL/5> Paljud treenerid suunavad tegelema pilatesega või joogaga. <T><VASES> Nad </VASES></T> mõlemad annavad rohkem enesekindlust juurde, aitavad kaasa mõtlemisele, rahustavad ning peletavad mured eemale.

<KOK/GÜ/2007/VIL/64> Muutugu ajad ja kombed palju tahes, inimene jääb inimeseks ja usun, et <T><VASES> kõik </VASES></T>, mis puudutab tundeid, ei puudunud ka neilt.

<KOK/GÜ/2007/VIL/2> Enamus linnalapsi on väga targad tänu paramatele koolidele linnas ning <T><VASES> kel </VASES></T> on olemas oma iidol ehk tähtis suurkuju, kellega sarnaneda.

Ebaselgelt viitav sõna

Kirjutaja viitab ebamääraselt minigile entiteedile, kuid ei ole täpselt aru saada, millele ta viidata soovib.

Tekst	<T><T/>	ei ole aru saada, kellele/millele sõna viitab	<T><?VIIT></?VIIT></T>

<KOK/GÜ/2007/VIL/17> Kõiki inimesi ei saa aga liigses tarbimises süüdistada, sest on inimesi, kellel pole <T><?VIIT> selleks </?VIIT></T> võimalusi ning neid, kes on <T><?VIIT> sellega </?VIIT></T> kaasnevatest probleemidest teadlikud.

<KOK/ G/2007/VIL/7> Ei saanud nad koolis ega koosolekutel käia. Nüüdsel ajal aga <T><?VIIT> seda </?VIIT></T> ei eksisteeri.

<KOK/GÜ/2007/VIL/34> Kui pöörduda nüüd tagasi minu eespool esitatud küsimuse juurde, et miks siiski inimesed eelistavad maaelule linnaelu, oleksid peamisteks põhjusteks paremad koolid, töökohad ja suuremad võimalused vaba aja veetmiseks. Samuti tuleb arvestada ka tõsiasja, et paljude inimeste jaoks on esmatähtsaks aspektiks just inimeste arvamus <T><?VIIT> nendest </?VIIT></T>.

<KOK/GÜ/2007/VIL/24> Ka lapsed võivad tõmbuda isast eemale, kuna <T><?VIIT> neil </?VIIT></T> pole ühel hetkel enam millestki rääkida.

Kahetimõistetavus

Tekst	<T><T/>	kahetimõistetavus	<T><KAH></KAH></T>

kahetähenduslik lause, väljend

<KOK/GÜ/2005/TRT/89/M/16> <T><KAH>
kui õnn ka tuleb, ei pruugi see olla pikk </KAH></T>.

<KOK G/2007/VIL/9> Noorena oli ta pärit jõukast perest, kogu elu oli ta saanud kõike, mida vaid hing ihaldas: autod, luksusliku korteri, <T><KAH> ostetud ülikooli </KAH></T>. lõputunnistuse, töökoha, kus mehe isa maksis tööandjale peale rohkem, kui mees palka sai.

sõnajärjest tingitud kahetimõistetavus

<KOK/GÜ/2007/VIL/24> Pikapeale kuhjuvad mured emal üle pea ning ta muutub närviliseks, mis omakorda mõjutab pere nooremaid liikmeid, kelle peale <T><KAH> valatakse kogu viha, sest meest pole kodus </KAH></T>..

<KOK/GÜ/2007/VIL/03> <T><KAH> Armunult oma peegelpilti imetledes lahkub Narkissosest allika ääres eluvaim </KAH></T>.

Sidususviga

Tekst	<T></T>	sidususviga	<T><SID></SID></T>

<KOK/GÜ/2007/VIL/3> <T><SID> Salastatud info teadasaamine seda avalikuse eest varjata sooviva inimese, muudab ta meie teenriks. </SID></T>

<KOK/ G/2007/VIL/7> <T><SID>On ju mehi, kellel on valitud sellised elukutsed, kus eelkõige leitakse, et naine peaks seda tegema. </SID></T>

<KOK/GÜ/2007/VIL/21><T><SID>Nii lülitas nii mõnigi oma mõtlemise välja ja usaldas jäägitult seda, mida päevast päeva kõrvuskumises. </SID></T>

<KOK/GÜ/2007/VIL/2> <T><SID>Paljud linnalapsed on ka mitmesuguste kompleksidega, mida on eriti noormeestel </SID></T>

Sõnastusviga

Ebaõnnestunud sõnastuse tõttu mõistetamatu, segane lause

Tekst	<T></T>	sõnastusviga	<T><SÕN></SÕN></T>

<KOK/GÜ/2007/VIL/5> <T><SÕN>
Sportlased ohverdavad väga palju aega treeningutele, kuid see toob ainult positiivseid tulemusi </SÕN></T>

<KOK/GÜ/2008/VIL/63> <T><SÕN>
Kuid mis siis, kui meile kõigile on saatus ette määratud ja me teeme vale valiku? </SÕN></T>

Sisuviga

Tühisõnalised laused, arusaamatuks jääva sisuga laused, faktivead.

Tekst	<T></T>	sisuviga	<T><SISUL></SISUL></T>

<KOK/GÜ/2007/VIL/49> <T><SISUL>
Tagajärjed on enamasti paratamatud </SISUL></T>

<KOK/GÜ/2007/VIL/33> <T><SISUL> Siis oleksid kõik vaesed haritud, asotsiaalid </SISUL></T>

<KOK/GÜ/2007/VIL/39> <T><SISUL> Isiksus jääb mõjutajaks vaid negatiivse alatooniga </SISUL></T>.

6.3.5. Märgendatud vigade parandamine

Õppija tekstides esinevate vigade märgendamisega koos või veatähiste sisestamise järel on kasulik vead ka parandada. Vigade parandamine on informatiivne nii selle poolest, et veatüüpide uurimisel on sellisel juhul võimalik arvesse võtta ka pakutud parandusvariante, kui ka seetõttu, et parandustega teksti on võimalik kasutada veaanalüsaatori arendamiseks.

Vigade märgendamine käsitsi, ilma sobivat programmi kasutamata on aeganõudev, kuid teatud aja möödudes päris hästi teostatav. Paranduste käsitsi sisestamine on väga veaohtlik, vajab erilist tähelepanu ning märgendites orienteerumist.

Järgnevalt kirjeldatakse vea parandamisel kasutatavaid põhimõtteid. Esmalt neid juhtumeid, mil lauses vajab parandamist üks keelend ning seejärel keerulisemaid juhtumeid.

Õppijakeele lause on järgmine
<KOK/GÜ/2007/VIL/25>
Õnneks pole sport ainult oma ala tippude harastus.

Esmalt tähistatakse veapiir vastava märgendiga:
<KOK/GÜ/2007/VIL/25> **Õnneks pole sport ainult oma ala tippude** <O><OK> **harastus** </OK></O>.

Seejärel lisatakse märgendi esimese poole järele veaparandus, mille märgend koosneb paranduse tähisest PAR + veatüübi lühendist, esitatud näites on veaparanduse tähiseks <PAROK>. Veaparandus lõpeb veaparanduse lõppu tähistava märgendiga </PAROK>. Sellele järgneb õppijakeele vigane vorm ning vigase vormi lõpetamist tähistav märgend.

Esitatud näite veaparandus
<KOK/GÜ/2007/VIL/25> **Õnneks pole sport ainult oma ala tippude** <O><OK> <PAROK> harrastus </PAROK> **harastus** </OK></O>.

Kui viga on võimalik parandada mitmel viisil, esitatakse mõlemad parandamisvõimalused.

Näiteks lause

Maailmas elab palju mitte lihasööjaid.

Lausele lisatakse veamärgend

Maailmas elab palju <N><PF> **mitte lihasööjaid** </PF></N>.

Lisatakse parandusvõimalused. Kuna selle lause parandamiseks on kaks võimalust, siis tähistatakse 1. neist märgendiga<1></1> ja teine märgendiga <2></2>.

Esitatud näite puhul näeks tekst, millesse on mõlemad parandusvõimalused sisestatud, välja jägmine

Maailmas elab palju <N><PF><PARPF> <1> inimesi, kes ei söö liha </1> <2> taimetoitlasi </2></PARPF> **mitte lihasööjaid** </PF></N>.

Järgnevalt demonstreeritakse lauset, milles on mitu erinevat veatüüpi.

<KOK/GÜ/2007/VIL/5> **Liikumisharrastuse levik eeldab aga kaasaegseid spordikomplekse ja treeningvahendeid ning seda adekvaatselt elanike vajadustele.**

Lauses võib eristada kolme viga:

1) nominalisatsioon – **liikumisharrastuse levik eeldab**
2) sõnatähenduse viga – **adekvaatselt**
3) parafraseerimine – **ning seda adekvaatselt elanike vajadustele**

Teksti lisatakse esimene veamärgend

<S><NOM> **Liikumisharrastuste levik eeldab** </NOM></S>

Eesmise märgendi <S><NOM> järele lisatakse selle vea paranduse tähis <PARNOM>, parandus ning paranduse lõppu tähistav märgend </PARNOM>. Paranduse lõppu tähistava märgendi järel on õppija vigane keelend ning selle järel veapiiri lõppu tähistav märgend </NOM></S>.

<S><NOM><PARNOM> Liikumisharrastuse levitamiseks on vaja </PARNOM> **Liikumisharrastuste levik eeldab** </NOM></S>

Teksti lisatakse teine veamärgend
<S><TAH> **adekvaatselt** </TAH>
ning selle vea parandus
<S><TAH><PARTAH> vastavalt </PARTAH> **adekvaatselt** </TAH>

Teksti lisatakse kolmas veamärgend
<N><PF> **ning seda adekvaatselt elanike vajadustele** </PF></N>.

ning selle vea parandus
<N><PF><PARPF> mis vastaksid elanike vajadustele </PARPF> **ning seda adekvaatselt elanike vajadustele** </PF></N>.

Märgendeid liites saame järgmiste tähistustega lause, mis sisaldab kõiki tekstis tehtud parandusi

<S><NOM><PARNOM> Liikumisharrastuse levitamiseks on vaja </PARNOM> **Liikumisharrastuste levik eeldab** </NOM></S> **aga kaasaegseid spordikomplekse ja treeningvahendeid** <N><PF><PARPF> mis vastaksid elanike vajadustele </PARPF> **ning seda** <S><TAH><PARTAH> vastavalt </PARTAH> **adekvaatselt** </TAH> **elanike vajadustele** </PF></N>.

Kui parandusvariante on mitu või vead kombineeruvad, kasutatakse parandusvariantide eristamiseks numbreid. Kõik ühe parandusvariandi juurde kuuluvad parandused on tähistatud sama numbriga. Märk # tähistab, et parandamiseks tuleb keelend tekstist välja jätta.

Õppijakeele lause: <KOK/GÜ/2007/VIL/5> **Varasematel perioodidel pühendati tervislikule eluviisile rohkem aega oma kiirest elutempost, kui praegu.**

Veamärgendiga võib tähistada

I parandusviis:

pühendati – vale ajavormi kasutamine

pühendati – vale tegumoe kasutamine

elutempo – tähendusviga

, – liigne kirjavahemärk

II parandusviis:

pühendati – vale ajavormi kasutamine

oma kiirest elutempost – liigne fraas

, – liigne kirjavahemärk

Varasematel perioodidel <X><SNJRG> <PARSNJRG> <1> on inimesed oma kiirest elust spordile rohkem aega pühendanud kui praegu </1> <2> on spordile rohkem aega pühendatud kui praegu </2> </PARSNJRG> <X> <VT><X><VKV><PARVKV><1> inimesed</1><PARVKV><PARVT> <1> on pühendanud </1> <2>on pühendatud</2></PARVT> **pühendati** <X></VKV></VT></X> **tervislikule eluviisile rohkem aega** <S><LIIG><PARLIIG><2>#</2></PARLIIG> **oma kiirest** <N><TAH><PARTAH> elust </PARTAH> **elutempost** </TAH></N> </LIIG></S> <I><LMARK><PARLMARK>#</PARLMARK> **,** </LMARK></I> **kui praegu** </SNJRG></X>.

Paranduste käsitsi sisestamine on väga veaohtlik, kuna ühel ajal tuleb arvesse võtta väga paljusid tegureid. Märgendussüsteemi efektiivseks kasutamiseks on vaja luua veamärgendite sisestamise programm, et märgendust automatiseerida ning märgendite sisestamisel tekkivaid vigu vältida.

6.4. Korpuse edasine arendustöö

Emakeeleõppija korpust on võimalik arendada nii kvalitatiivnselt kui kvantitatiivselt. Korpuse kvalitatiivse arendamise all mõtlen sobiva veaprogrammi kirjutamist, õppijate tekstide märgendamist ning analüüsimist. Eesmärgiks on märgendada iga kooliastme tekstidest vähemalt 25 000-sõnaline osa.

Kvantitatiivse arendamise all mõtlen tekstide kogumist. Praegu emakeelekorpuses sisalduvate tekstide põhjal on võimalik analüüsida gümnaasiumiõpilaste keelekasutust, kuid kogu ei ole tasakaalustatud ega kuigi mahukas. Korpus võiks pakkuda paremaid tekstide võrdlemise võimalusi. Emakeeleõppija korpus ei sisalda veel I ega II kooliastme õpilaste tekste, mis õppijakeele arengu jälgimise seisukohalt oleksid samuti olulised. Vaja on koguda ka materjali nn kestvusuuringute jaoks.

Eesti keele õppija kirjaliku korpuse võimalikeks koostisosadeks on kolm omavahel seotud allkorpust – lisaks emakeeleõppija korpuse tekstidele sisaldaks see ka eesti keelt teise keele või võõrkeelena õppijate võrdluskorpust ja sisendkeele tekstidest koostatud eeskujukorpust. Kõikide korpuse osade eesmärgid on erinevad, kuid ühine märgendussüsteem seob nad mahukaks *eesti keele õppija kirjalikuks korpuseks*.

Emakeeleõppija korpuse tekstid kogutakse emakeelt õppivate laste ja noorte kirjaliku keelekasutuse uurimiseks. Võrdluskorpuse tekstid kogutakse eesti keelt teise keelena või võõrkeelena õppivatelt lastelt ja noortelt. Mõlema allkorpuse tekstid kogutakse ühesugustelt alustelt. Õpilastekstid ei tohi olla tervikulatuses kättesaadavad ning töö autor ei tohi olla tuvastatav.

Metainformatsioonina lisatakse õppijate tekstidele

1. õpilase vanusegrupp;
    - I kooliaste 1.–3. klass;
    - II kooliaste 4.–6. klass;
    - III kooliaste 7.–9. klass;
    - gümnaasiumiõpilased 10.–12. klass.

2. õpilase emakeel
3. õpilase sugu;
4. kirjutaja asukoht (piisab ka määratlusest Lõuna-Eesti, Kesk-Eesti, Põhja-Eesti, Lääne-Eesti);
5. teksti kirjutamise aeg: aastaarv;
6. teksti tüüp;
7. teave teksti kirjutamise situatsiooni kohta.

Kolmas eesti keele õppija korpuse osa on eeskujukorpus, mille tekstid kogutakse keelesisendi uurimiseks ning õppijatele kohase eeskujukorpuse loomiseks. Eeskujutekstid jagatakse valdkonniti (näiteks loodus- ja reaalteaduslikud kirjutised; humanitaarteaduslikud kirjutised; sotsiaalteaduslikud kirjutised; ametkondliku suhtlemise tekstid; varia). Eeskujutekstideks sobivad populaarteaduslikud artiklid, teatmeteoste tekstid, õpikutekstid (erinevate õppeainete õpikutekstid), ajakirjandustekstid (ka lasteajakirjadest), ilukirjanduslikud tekstid; filmisubtiitrid, reklaamtekstid, laulutekstid jm.

Eeskujukorpuse tekstidele lisatakse metaandmetena

1. teksti teema
2. teksti laad (ajakirjandus, ilukirjandus, õpikutekst, populaarteaduslik tekst)
3. teksti pealkiri
4. väljaandmisaeg: aastaarv
5. autor

Eesti keele õppijate korpus on vajalik nii eesti keelt emakeelena kui ka teise või võõrkeelena õppijate keelekasutuse uurimiseks.

Kokkuvõte

Minu magistritöö käsitles eesti emakeelega õpilaste keelekasutust sisaldava korpuse loomise ja uurimisega seonduvat. Korpuse loomist käsitlev teoreetiline osa põhineb peamiselt Silviane Grangeri, Douglas Biberi ning Susan Hunstoni seisukohtadel. Õppijakorpuse analüüsimeetodi – veaanalüüsi kirjeldus tugineb suuresti Carl Jamesi, ning Rod Ellise seisukohtadele.

Töö peamiseks eesmärgiks oli luua emakeeleõppija korpus ning pakkuda välja selle edasise arendamise võimalusi. Eesmärgi saavutamiseks uurisin õppijakorpuse koostamise ja uurimise põhimõtteid, õppijakorpuse rakendamise võimalusi ning mujal maailmas loodud õppijakorpusi.

Ingliskeelseid õppijakeele korpuseid on maailmas koostatud alates eelmise sajandi 90. aastatest. Eestis on korpuspõhiselt uuritud väikelaste keelt ja mitte-eesti emakeelega täiskasvanute keelekasutust. Eesti keelt emakeelena õppivate koolilaste tekstikorpust seni loodud ei ole.

Töö teoreetilises osas tutvustati õppijakorpuse eripära ning selle koostamisega seotud üksikasju, mis suurte ning üldist keelekasutust analüüsivate korpuste loomisel ei ole määrava tähtsusega, kuid muutuvad oluliseks spetsiifilise eesmärgiga allkorpuste koostamisel. Korpuste loomist kirjeldati emakeeleõppija korpuse loomise vaatepunktist ning analüüsisiti teksti kogumise põhimõtteid tekstist, autorist, kirjutamise situatsioonist ja ülesandest lähtuvalt. Katseliselt loodud 336 389-sõnaline emakeeleõppija korpus sisaldab peamiselt gümnaasiumiõpilaste esseelaadseid tekste, mis õppijad on kirjutanud kooli- või võistlustööna. Kirjanditele on lisatud metaandmed teksti kirjutamise situatsiooni, kirjutamise aja, autori vanuse ja päritolu kohta.

Õppijakorpuse uurimise võimalustest tutuvustati nii märgendamata kui ka märgendatud korpuse analüüsi meetodeid (statistiline ja veaanalüüsi meetod) ning tehti õpilastekstide statistiline analüüs. Uuriti õppijate lausete keskmist pikkust ning sagedamini kasutatavaid sõnaliike: verbe, nimisõnu ja asesõnu. Tulemusi on võrreldud eesti kirjakeele sagedussõnastiku ning vahekeele korpuse andmetega.

Märgendatud õppijakorpuse analüüsivõimalustest on pikemalt kirjeldatud veaanalüüsi etappe ning magistritöö praktiline osa keskendub P. Corderi jaotuses veaanalüüsi kolmandale ehk vigade klassifitseerimise etapile.

Õppijatekstides sisalduvate vigade süsteemi loomiseks on parandatud 10 000 sõna suurune gümnaasiumiõpilaste kirjandite kogu. Õppijate vigadest ja kooligrammatika käsitlusviisist lähtuvalt on vead jagatud üldkategooriatesse: ortograafia, interpunktsioon, morfoloogia, süntaks, semantika, stiil ja tekst. Igaks juhuks on tekitatud ka kaheksas üldkategooria (*muu*), millesse kuuluvad need vead, mis eelpoolnimetatud üldkategooriatesse ei liigitu. Veaohtlikud keelendid (vt Erelt 2006, Kasik 2007), mis õppijate kirjandites ei esinenud, veagruppides ei kajastu.

Veakategooriad on jagatud veatüüpide järgi allkategooriateks, mille detailne jaotus võib täpsustuda korpuse arendustegevuse käigus. Iga veakategooria all on esitatud näiteid õppijate tekstidest.

Veaklassifikatsiooni alusel on võimalik tekste märgendada. Märgendamise eesmärgiks on teksti eripära ja õpilaste keelekasutuse uurimine, mitte hindamine, mistõttu tähistatakse ka keelendid, mis ei ole otseselt vead, kuid näitavad õppijakeeles teatud tendentsi või eelistust.

Vealiikide märgendamise aluseks on võetud prantsuse õppijakorpuste looja ja uurija Silviane Grangeri (2003) õppijakorpuse FRIDA veamärgendussüsteem. Emakeeleõppija korpuses on võimalik vigu analüüsida kahel tasandil: üldtasandil ja vealiigiti. Vajadusel võib tasandeid lisada. Loodud märgendite süsteem on aluseks TÜ eesti keelt võõrkeelena õppijate veakorpuse märgendussüsteemile.

Korpuse edasiseks arendamiseks märgendatakse kogutud tekste ning suurendatakse eesti keele õppijate tekstikogu: korpus peaks emakeeleõppijate tekstide kõrval sisaldama ka valikut sisendkeele tekstidest ning eesti keelt teise keele või võõrkeelena õppivate laste ja noorte keelekasutust.

Emakeeleõppija korpust saab rakendada õpetajakoolituses, korpuspõhise uuringu tulemusi on võimalik kasutada ainekavade ja õppematerjalide koostamisel.

Minu magistritöö juhtis tähelepanu eesti keele õppijate tekstide eesmärgipärase kogumise ning eesti keele õppija korpuse loomise vajadusele.

Corpus of Native Estonian Learners
Statistical Analysis and Error Tagging System

Résumé

The current Master's thesis is investigating the issues of compiling and researching a learner corpus of native Estonian schoolchildren.

The main objective of the thesis was to create a corpus of native Estonian schoolchildren and propose the possible ways for its further development. To obtain the goal, the principles of creating and researching learner corpora were studied as well as existing learner corpora in the other countries.

Learner corpora in English have been compiled since 1990s. In Estonia, corpus-based research has been done to investigate child language and non-Estonian speaking adults' language usage. Corpus containing the language data of native Estonian schoolchildren does not exist yet.

The first chapter of the Masters' thesis gives an overview of different types of corpora; their characteristics and issues about compiling a corpus. Theoretical principles of creating a corpus are based on Silviane Granger's and Douglas Biber's views. In the second chapter, principles of selecting texts for learner corpus are presented. The third chapter deals with using corpora in language teaching. Corpus-based research is introduced in the fourth chapter of the thesis. In the fifth chapter, error analysis as a widely used approach to study learner texts is introduced. A description of error analysis mainly relies on Carl James's, Pit Corder's and Rod Ellis's studies. The sixth chapter contains a description of compiled corpus of native Estonian learners, its statistical analysis and the error tagging system.

Illustrative materials can be found in appendixes.

The experimentally created learner corpus contains 336 389 words and consists mainly of essays written by high school students. Metadata added make it possible to analyze learner language from the point of writing situation and time as well as the age and origin of the author.

The statistical analysis of the learner corpus was carried out. The average length of sentences and most frequent word classes (verbs, nouns and pronouns) were analyzed. The results were compared with the data of Estonian word frequency (Kaalep & Muischnek 2002) and Estonian Interlanguage Corpus (*Eesti vahekeelekorpus*).

To work out learners' error system, texts in amount of 10 000 words were corrected. Relying on school grammar and P. Corder's research, errors were classified into general categories: spelling, punctuation, morphology, syntax, semantics, style and text. An additional category ("other") was created for errors not possible to be classified under any of the mentioned categories. Potential errors (see Erelt 2006, Kasik 2007) that were not identified in learners' texts are not incorporated in error groups. Error categories are divided into subcategories, which may be specified in details while further development of the corpus. Errors classified are illustrated with the examples of learner texts.

Error tagging system created for the learner corpus is based on the system used in French learner corpus FRIDA (see Silviane Granger 2003, 2006, 2007), enabling error analyses according to either general level or error groups. If needed, additional error levels can be created.

As a further development of the created corpus, an extensive learner corpus could be compiled containing native Estonian students' texts as well as non-Estonian students' Estonian language usage and selection of input texts. Corpus-based research can be used to analyze learner language, also to improve syllabus and study materials design. Native Estonian learner's corpus can also be used in student-teaching programs and teachers' in-service training.

The current Masters' thesis calls attention to the need for systematic gathering of the learner texts and creating a corpus of the Estonian language learner.

Kasutatud materjalid

Argus, Reili 2007. Eesti lastekeelekorpuse morfoloogilisest märgendamisest. – Tallinna ülikooli keelekorpuse optimaalsus, töötlemine ja kasutamine. Toimetanud Pille Eslon. Tallinn: Tallinna Ülikooli kirjastus, lk 65–86.

Asser et al. = *Asser, Hiie; Heiki-Jaan Kaalep; Siret Linnas; Jaan Mikk; Kadri Muischnek; Merje Songe; Heli Uibo* 2004. Õpikute keerukuse analüüs arvutitel. – Toimiv keel II Töid rakenduslingvistika alalt. Tallinn: TPÜ kirjastus, lk 72–84.

Barlow, Michael 2005. Computer-based analyses of learner language. – Analysing Learner Language. Barkuizen, Ellis. Oxford: Oxford University Press.

Bernardini, Silvia 2007. Corpora in the classroom: an overview and some reflections on future developments. – Corpus Linguistics. Critical Concepts in Linguistics *vol* VI. *Ed by* Wolfgang Teubert & Ramesh Krishnamurthy. Routledge, lk 102–123.

Biber Douglas, Susan Conrad, Randi Reppen 2006. Corpus Linguistics. Investigating Language Structure and Use. Cambridge: United Kingdom at the University Press.

Bick, Eckhard 2004. Live Use of Corpus Data and Corpus Annotation Tools in CALL: Some New Developments in VISL. http://beta.visl.sdu.dk/pdf/corpus_and_CALL_form.pdf. 2.05.2008.

Bowker, Lynne, Jennifer Pearson 2002. Working with Specializd Language. A Practical Guide to Using Corpora. Bury St Edmunds, Suffolk, Great Britain: St Edmundbury Press.

Conrad, Susan 2007. Corpus Linguistics. Critical Concepts in Linguistics *vol* VI. *Ed by* Wolfgang Teubert & Ramesh Krishnamurthy. Routledge, lk 55–66.

Concept to classroom. A series of workshops.
http://www.thirteen.org/edonline/concept2class/constructivism/index_sub2.html3.01.2008

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

Diaz-Negrillo, Ana, Jesus Fernandez-Dominiguez 2006. Error tagging Systems for Learner Corpora. – RESLA nr 19, lk 83–102.

Graham Davies, Lamy Marie-Noelle, Mortensen H. Jorgen 2008. Using concordance programs in the Modern Foreign languages classroom. – Information and Communications Technology for Language Teachers (ICT4LT Module 2.4.) http://www.ict4lt.org/en/index.htm. 23.04.08.

Gerassimenko, Olga 2008. Eravestlus.

Ehala, Martin 1997. Kommentaarid ”Eesti keele struktuuri” juurde. Tallinn: Künnimees.

Ehala, Martin 1998. Eesti kirjakeel. Tallinn: Künnimees.

Ehala et.al = *Ehala, Martin, Tiina Veismann, Aili Kiin, Astrid Aus* 2005. Noor keelekasutaja. Põhikooli vanema astme eesti keele metoodika ja õpetajaraamat. Tallinn: Künnimees.

EKK = *Erelt, Mati, Tiiu Erelt, Kristina Ross* 1997. Eesti keele käsiraamat. Tallinn: Eesti Keele Sihtasutus.

EKR = Euroopa keeleõppe raamdokument: õppimine, õpetamine ja hindamine 2007. Haridus- ja Teadusministeerium. Tartu.

Ellis Rod, Gary Barkhuizen 2005. Analysing Learner Language. Oxford: Oxford University Press.

Erelt, Mati 2006. Lause õigekeelsus. Tartu: Bookmill.

Eslon, Pille 2007. Õppijakeele korpused ja keeleõpe. Tallinna ülikooli keelekorpuse optimaalsus, töötlemine ja kasutamine. Toimetanud Pille Eslon. Tallinn: Tallinna ülikooli kirjastus, lk 87–120.

Eslon Pille, Helena Metslang 2007. Õppijakeel ja eesti vahekeele korpus. – Eesti rakenduslingvistika aastaraamat. Tallinn: Eesti Keele Sihtasutus, lk 99–116.

Granger, Sylviane 2003. A Promising Synergy. – Error-tagged Learner Corpora and CALL: CALCIO 20 (3), lk 465–480.

Granger Sylviane 2006. Centre for English Corpus Linguistics. http://www.fltr.ucl.ac.be/FLTR/GERM/ETAN/CECL/research%20learner%20corpora.html. 10.05.2008.

Granger, Sylviane 2007. A birds-eye view of learner corpus research. – Corpus Linguistics. Critical Concepts in Linguistics *vol* VI. *Ed by* Wolfgang Teubert & Ramesh Krishnamurthy. Routledge, lk 23–44.

Habicht, Külli 2006. Eesti kirjakeele hetkeseis: tähelepanekuid ja probleeme. http://www.fl.ut.ee/orb.aw/class=file/action=preview/id=118448/kyllihabicht.doc. 18.04. 2008.

Habicht, Külli 2007. Kuidas vältida vigu eksamikirjandis. TÜ täienduskoolituskursuse materjalid.

Habicht, Külli 2008. Kirjandieelset keelenõu õpetajatele. TÜ täienduskoolituskursuse materjalid.

Habicht, Külli, Pille Penjam 2006. Kaassõna keeleuurija ja -kasutaja käsituses. – Emakeele Seltsi aastaraamat nr 52. Tallinn: Teaduste Akadeemia kirjastus, lk 51–68.

Hennoste Tiit, Kadri Muischnek 2000. Eesti kirjakeele korpuse tekstide valiku ja märgendamise põhimõtted ning kahe allkeele võrdluse katse. – Arvutuslingvistikalt inimesele Tartu: Tartu ülikooli üldkeeleteaduse õppetooli toimetised 1, lk 183–217.

Hennoste, Märt 2008. Eravestlus.

Hunston, Susan 2002. Corpora in applied Linguistics. Cambridge: Cambridge University Press.

James Carl 1998. Errors in Language Learning and Use. Exploring error analysis. London and New York: Longman.

Johns, Tim 2007. Should you be persuaded: two samples of data-driven learning materials. – Corpus Linguistics. Critical Concepts in Linguistics *vol* VI. *Ed by* Wolfgang Teubert & Ramesh Krishnamurthy. Routledge, lk 3–23.

Jonassen, David H. 1994. Thinking technology: Toward a constructivist design model. – Educational Tecnology, March–April, lk 34–37.

Jürine, Anni 2007. Eesti kaassõnaühendite norminguvastasest kokkukirjutamisest grammatiseerumisteooria taustal. Bakalaureusetöö käsikiri TÜ eesti keele osakonnas.

Jüssi, Elise. 2005. Laadivahelduse taandumine eesti kirjakeeles õpilaste keelekasutuse näitel. Bakalaureusetöö käsikiri Tartu Ülikooli eesti keele osakonnas.

Kaalep, Heiki-Jaan 1999. Eesti keele ressursside loomine ja kasutamine keeletehnoloogilises arendustöös. Diss. Tartu: Tartu Ülikooli kirjastus.

Kaalep, Heiki-Jaan, Kadri Muischnek 1999. Eesti kirjakeele sagedussõnastik. Tartu: Tartu Ülikooli Kirjastus.

Kaalep, Heiki-Jaan, Toomas Vaino 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. Arvutuslingvistikalt inimesele. Tartu: Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1, lk 87–100.

Kaalep, *Heiki-Jaan* 2008. Eravestlus.

Kasik et al. = *Kasik, Reet, Mati Erelt, Tiiu Erelt* 2007. Eesti keele väljendusõpetus kõrgkoolidele. Tartu.

Kerge, Krista 2002. Aja- ja ilukirjandusteksti süntaktilise keerukuse dünaamika XX sajandil. TPÜ eesti keele osakonna veebitoimetised. Lingvistika 1 http://digar.nlib.ee/otsing/stream?pid=nlibdigar:11198&file_id=VAATAMINE_3_1. 10.05.2008.

Kerge, Krista 2005. Nominalisatsion ja nominaalstiil. – Oma keel nr 2, Emakeele Selts, lk 37–42.

Krüger, Anja, Simon Hamilton 1997. RECALL: individual language tutoring through intelligent error diagnosis. – ReCall vol 9 nr 2 1997 lk 51–58.

Kuresoo, Uno 1990. Keelevead õpetajate ja keeleteadlaste hinnanguis. – Keel ja Kirjandus nr 8, lk 462–470.

Lepajõe, Kersti 2006. Õpilaste registri- ja stiilitajust riigieksamikirjandite põhjal. – Keele ehe. Tartu: Brookmill, lk 141–153.

Lamy et al. = *Lamy, Marie-Noelle, Hans Jorgen Klarskov Mortensen, Graham Davies* 2007. Using concordance programs in the modern foreign languages classroom. – ICT4LT Module 2.4 www.ict4lt.org/en/en_mod2-4.htm 11.04.2007.

Lemnitzer Lothar, Heike Zinsmeister 2006. Korpuslinguistik. Eine Einführung. Gunten Narr Verlag Tübingen.

Madissoo Maia 1998. Sõnastusharjutusi eksamikirjandiks valmistujatele. Tallinn: Koolibri.

Madissoo, Maia 2002. Keele ja tekstiharjutusi abiturientidele. Tartu: Künnimees.

McEnery Tony, Andrew Wilson 1997. Corpus Linguistics. Edinburgh University Press.

McEnery et al. = *McEnery, Tony, Andrew Wilson, Paul Barker* 1997. Teaching grammar again after twenty years: corpus-based help for teaching grammar. ReCall vol 9 nr 2 nov, lk 8–17.

Meier Heidi 2003. Essee allkeelte võrdluses. – Tekstid ja taustad II. Tekstianalüüsi vaatepunkte. Tartu: Tartu Ülikooli kirjastus, lk 116-135.

Meunier, Fanny 2007. The pedagogical value of native and learner corpora in EFL grammar teaching. – Corpus Linguistics. Critical consepts in Linguistics *vol* VI. *Ed*. Wolfgang Teubert & Ramesch Krishnamurthy, lk 23–44 Routledge.

Muischnek, Kadri 2006a. Korpuslingvistika loengukursus.

Muischnek, Kadri 2006b. Verbi ja noomeni püsiühendid eesti keeles. Diss. Tartu: Tartu Ülikooli kirjastus, lk 9–45.

Muischnek, Kadri 1998. Korpuslingvistika. – Keel ja Kirjandus nr 1, lk 8–13.

Muischnek et al = *Muischnek, Kadri, Heili Orav, Heiki-Jaan Kaalep, Haldur Õim*. Eesti keele tehnoloogilised ressurssid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Tallinn: Eesti Keele Sihtasutus. www.hm.ee/index.php?popup=download&id=3993.

Müürisep, Kaili 2008. Eravestlus.

Nesselhauf, Nadja 2005. Collocations in a Learner Corpus. Amsterdam/Philadelphia. John Benjamins Publishing Company.

O´Keeffe et al. = *O´Keeffe, Anne, Michael McCarthy, Ronald Carter* 2007. From Corpus to classroom. Language use and language teaching. Cambridge: Cambridge University Press.

Pajusalu, Renate 1999. Deiktikud eesti keeles. Diss. Tartu: Tartu Ülikooli kirjastus.

Pajusalu et al. *Pajusalu, Renate, Ilona Tragel, Ann Veismann, Maigi Vija* 2004. Tuumsõnade semantikat ja pragmaatikat. Tartu: Tartu Ülikooli kirjastus.

Pajusalu, Renate 2004. Keeleõppija tähendusviga ja selle hindamine. – Toimiv keel II. Töid rakenduslingvistika alalt. Tallinna: Tallinna Pedagoogikaülikooli kirjastus, lk 218-236.

Paomees, Eve. Suulise kõne ilmingud 2000. aasta riigieksamikirjandites. Bakalaureusetöö. Käsikiri eesti keele osakonnas.

Pool, Raili 2007. Eesti keele teise keelena omandamise seaduspärasusi täis- ja osasihitise näitel. Diss. Tartu: Tartu Ülikooli kirjastus.

Pool, Raili, Elle Vaimann 2005. Vead kõrgtasemel eesti keele kõnelejate kirjalikus keelekasutuses. – Eesti rakenduslingvistika ühingu aastaraamat 1 (2004). Tallinn: Eesti Keele Sihtasutus., lk 115–137.

Pravec, Norma. Survey of learner corpora. ICAME Journal no 26, lk 81–114.

Probleemõppe lähtekohad. Konstruktivism. Sotsiaalkonstruktivistlik õppimiskäsitus. www.htk.tlu.ee. 15.12.2007.

Ranalli, James M. 2003. ELT coursebooks in the age of corpus linguistics: constraints and possibilities, lk 2–19. www.cels.bham.ac.uk/resources/essays/Ranally6.pdf 10.02.2008.

Römer, Ute 2004. Comaring real and ideal learner language input: The use of an EFL textbook corpus in corpus linguistics and language teaching. – Corpora and language

Learners. Edited by G. Aston, S. Bernardini, D. Stewart. Amsterdam. Philadelphia: J. Benjamins, 150–168 .

Rätsep, Huno 1978. Eesti keele lihtlausete tüübid. Tallinn: Valgus.

Peter Siemen, Anke Lüdeling, Frank Henrik Müller 2006. Falko – ein fehlerannotiertes Lernerkorpus des Deutschen. – Proceedings of Konvens, Konstanz.

Sinclair, John 1995. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Sinclair, John 2004. How to use Corpora in language Teaching. Amsterdam/Philadelphia: John Benjamins Publisching Company.

Tribble, Cris; Glyn Jones 1997. Concordances in the Classroom. A Source Book for Teachers. UK: Longman Group .

Tribble, Cris 1997. Improvising corpora for ELT: quick-and-dirty ways of developing corpora for language teaching. A paper presented at the First international conference: Practical Applications in Language Corpora. www.ctribble.co.uk. 12.02.2008.

Tribble, Cris 2008. Using concordance programs in the Modern Foreign languages classroom. – Information and Communications Technology for Language Teachers (ICT4LT Module 2.4.) http://www.ict4lt.org/en/index.htm. 23.04.08.

Tono, Yukio 2002. Learner corpora: design, development and applications. http://ucrel.lancs.ac.uk/publications/CL2003/papers/tono.pdf, 20.02.2008.

Uibo, Heli, Eckhard Bick 2005. Treebank-based research and e-learning of Estonian syntax. – The Second Baltic Conference on Human language Technologies. Tallinn: Tallinna Raamatutrükikoda, lk 195–200.

Uusen, Anne 2006. Põhikooli I ja II astme õpilaste kirjutamisoskus. Tallinn: Tallinna Ülikooli kirjastus.

Valmis, Aavo, Valmis Lembetar 2001. Õigekeelsuse käsiraamat gümnaasiumile. Tallinn: TEA.

Vija Maigi 2007. Pronoomenid lapsekeeles: mõnda *mina* ja *sina* omandamisest. – Eesti rakenduslingvistika aastaraamat. Tallinn: Eesti keele Sihtasutus, lk 373–383.

Villandi, Leo 1966. Sõnastusvigade liigid ja ravi. Tallinn: Valgus.

Vill Taimi 2001. Viga eesti keeles eesti kooliõpilase hinnangul Bakalaureusetöö käsikiri eesti keele osakonnas.

Willis, Dave, Jane Willis 2007. Designing task-based syllabus. – Doing Task-based teaching. Oxford: Oxford University Press, lk 177–198.

Willis, Dave 1991. Collins Cobuild Students grammar. Classroom edition. Helping leaners with real English. Harper Cullins Publishers.

Kaudviited

Corder, Pit 1974. Error Analysis. – The Edinburgh Course in Applied Linguistics, Vol 3. Edited by J. Alled, Pit Corder. Oxford: Oxford UniversityPress. Viidatud Ellis, Barkhuizen 2005. Analysing Learner Language. Oxford University Press, lk 57 kaudu.

Leech, Geoffery 1997. Introducing Corpus Annotation. – Corpus Annotation. Linguistic Information from Computer Text Corpora, Garside, Roger, Geoffrey Leech, Tony McEnery. London/ New York: Longman, lk 1–18. viidatud Lemnitzer, Lothar, Heike Zinmeister 2006. Korpuslinguistik. Eine Einführung. Gunten Narr Verlag Tübingen, lk 96 kaudu.

Lyons, J. 1977. Semantics. Cambridge University Press, Cambridge. Viidatud James, Carl 1998. Errors in language Learning and Use. London and New York: Longman lk 69 kaudu.

Sökmen, L. 1997. Current Trends in Teaching Second Language Vocabulary. – Vocabulary, Description, Acquisition and Pedagogy. Toimet N. Schmidt M. McCarthy. Cambridge: Cambridge University Press. Viidatud Granger, Silviane 2007. A birds-eye view of learner corpus research. – Corpus Linguistics. Critical Concepts in Linguistics *vol* VI. *Ed by* Wolfgang Teubert & Ramesh Krishnamurthy. Routledge, lk 62 kaudu.

Lisad

Lisa 1. Näide korpuse kasutamist emakeeletunnis

Kõikides järgnevates eesti kirjakeele korpusest võetud lausetes on
üks sõna asendatud sõnaga “tort”.
Mis sõna on lausetest välja võetud?
Taasta laused – kirjuta praeguse asemele õige sõna õiges vormis!!

Isegi ajalehes kirjutati sellest **tordist.**
Seekord oli kultuurimajas mingi **tort**.
Kuulutasin, et pääs **tordile** on kõigile vaba.
Nagu oleks ta viibinud tõesti **tordil**.
Esimene plaat on Bachi Brandenburgi **tordid**.
Tema korraldatavat "Pepsi Planeedi" nimelist **tort**tuuri saatis pidev vihmasadu.
Pidime enne **torti** tema poissmehe-elamises kokku saama.
Noorena olin ma suur muusikaarmastaja ja käisin hoolega **tortidel**.
Sellega loodame kiiresti korvata **tordi** kahjumi.
Nüüd saab seal **tortegi** korraldada.

Lisa 2. Näide korpuse kasutamisest emakeeletunnis

Esitatud on kümme näitelauset ajakirjanduskorpusest.
Loe lauseid ja otsusta, mida need laused illustreerivad.
Missugune lause erineb teistest ja mille poolest? Selgita!

☐ 1. Nimelt on see sõõrjas moodustis kolmekorruse**line** kuid ainsagi trepita.

☐ 2. Saadikurühma juhib viieliikme**line** kaasesimeeste kogu, mis vastab rahvarinde eesseisusele.

☐ 3. Põldudevahe**line** tee on sile nagu asfalt.

☐ 4. Pikaaja**line** lahusolek sünnimaa jaherõskest kliimast ei mõjunud noorukile hästi: Westholmis kutsusid klassivennad teda Inspektoriks - nii sageli tuli puududa.

☐ 5. Meie 30-liikme**line** poiste ja tüdrukute grupp koosneb 11. ning 12. klassi õpilastest, kes käivad kas inglise keele eriklassis või spordikallakuga klassis.

☐ 6. Kõige värskemana on meeles kolme liiduvabariigi vahe**line** Balti tuur, mis algas 27. mail Vilniuses ja lõppes 8. juunil Tartus.

☐ 7. Ühele võluvits külluslikku homsesse, teisele ränk töö, kolmandale roosaprilli**line** tulevik.

☐ 8. Suhteliselt väikeseliikme**line** trupp eesotsas kunstilise juhi ja korraldajadirektori Juri Šiškiniga on juba mitu lastetükki lavale toonud.

☐ 9. Imikule lisaks oli isal oma viiesajarea**line** poeem kaasas .

☐ 10. Sellemõtte**line** tekst on raiutud ka tema mälestuskivile .

Lisa 3. Näide korpuse kasutamisest sõnavara õpetamisel

Eesti kirjakeele üldkorpusest leitud näited selgitavad paronüümide *pitser* ja *pitsat* kasutust.

Üks lause esitatud 10 näitelause seas on väär.

Missugune?

☐ 1. Pärma käest saadud konservatooriumi **pitsat**iga tunnistus taskus, rändasin läbi kogu Põhjaranniku.

☐ 2. Riigiasutuste ja -organite , seega ka notarite **pitsat**ite jäljed peavad tulevikus sisaldama tunnust , mis võimaldab seda eristada kehtetuks tunnistatud või tulevikus registreeritavatest sama tekstiga **pitsat**ite jäljest .

☐ 3. Eesti Pank on juriidiline isik , tal on oma põhikiri, **pitsat**, vapp ja muu seadusega lubatud atribuutika.

☐ 4. Nimelt ei andnud **pitsat** alguses täisväärtuslikku jäljendit.

☐ 5. Ta leidis siit ühe Haapsalu toomhärra XIII sajandist pärit **pitsat**i, münte ja hõbekangikese .

☐ 6. See on uus säte selleks, et **pitser** oleks usaldusväärne .

☐ 7. Litsentsil peab olema all **pitser** ja kindel allkiri , Riigikogul tuleb siis luua selle komisjoni juurde mingi täitevvõimuorgan .

☐ 8. Ent oma selge **pitser**i vajutas tollele üsna juhuslikule ning eklektilisele raamatujadale paraku keskpärasus .

☐ 9. Tunnistuse " põhimõte tähendab propagandistlikule teatele autoriteedi ametliku **pitser**i andmist .

☐ 10. Riigina eksisteerimine paneb **pitser**i samuti referendumist osavõtjate ringi piiritlemisele.

Lisa 4. Ülekasutatud sõnad õppijakorpuses

(Ellis Barkhuizen 2005 : 352)

Word	Percentage Difference	Log Likelihood
is	1.22%	2360.12
people	0.43%	1998.98
we	0.56%	1983.56
this	0.53%	1302.74
will	0.40%	838.57
our	0.21%	811.08
they	0.37%	693.85
are	0.41%	677.17
moreover	0.05%	587.99
be	0.42%	577.98
not	0.37%	574.91
also	0.21%	572.66
can	0.23%	560.46
nowadays	0.04%	471.06
because	0.16%	452.60
think	0.11%	442.59
of	0.77%	419.51
to	0.70%	397.25
their	0.25%	378.56
it	0.36%	365.15
these	0.12%	328.59
indeed	0.06%	325.67
does	0.09%	319.61
that	0.38%	316.85

Lisa 5. Kõrgtasemel keeleõppijate vealiigid

Kõrgtasemel eesti keele õppijate kirjalikes töödes esinenud keelevigade liigitus (Pool 2007; Pool, Vaimann 2005).

vigade liigid	**esinemise arv**	**esiemise protsent**
sõnajärg	14	13,9
sõnavalik	7	6,9
täis- ja osasihitis	1	1
sõnastus	10	9,9
vormimoodustus	2	2,0
liigne sõna	1	1.0
ajavorm	1	1,0
käände valik	1	1,0
kokku- ja lahkukirjutamine	16	15,9
ühildumine	3	3,0
puuduv sõna	4	4,0
sõnakordus	2	2,0
poolik lause	4	4,0
mehhaanilised vead		
vale täht	5	5,0
liigne täht	2	2,0
muud juhtumid		
kaheti analüüsitavad vead	2	2,0
eri arvamus	5	5,0
interpunktsioon	21	20,8
kokku	101	100,0

Lisa 6. Gümnaasiumiõpilaste sagedamate keelevigade liigitus

Sagedamad keelevead gümnaasiumiõpilaste lõpukirjandite põhjal (2007–2008) (Habicht 2007, 2008).

ortograafiavead	häälikuortograafia võõrsõnade ortograafia kokku- ja lahkukirjutamine suure ja väikese algustähe vead
interpunktsioonivead	koma koondlauses täiendi koma määruse koma lisandi koma koma liitlauses lauselühendi koma võrdlussõnade komastamine
vormivead	käändsõnavead omadussõnavead verbivead asesõnavead
lausestusvead	rektsioon sõnajärg väärseosed *läbi* vahendi tähenduses, *omab*-konstruktsiooniga liialdamine
stiilivead	argikeelsus, moesõnad, lekseemi mittetundmisest tingitud sõnavalikuvead, paronüümia viitesuhted *sina*-stiil saama tulevik ajavormide kasutamine läbisegi hakitud stiil

Lisa 7. Sõnastusvead Leo Villandi järgi (1966)

vea liik	
sõnade tähendusvead	põhitähenduse ja tähendusvarjundi mittetundmine sõnade ärasegamine kõlalise sarnasuse alusel
kordus ja paljusõnalisus	pleonasm (paljusõnalisus) sõnatüve põhjendamatu kordamine mõtete põhjendamatu kordamine pikk, ülekoormatud lause tautoloogia samakõlaliste häälikute põhjendamatu kordamine (heakõlavead)
seosevead	mõtete omavaheline seostamatus ja järsud üleminekud vajaliku sõna puudumine ühtekuulumatute mõistete seostamine sõnade vale seostamine lauseosade (kõrvallause, lauselühendi) vale seostamine poolikud laused põhisõnast lahutatud omadussõnaline täiend piiritlemata asesõna tsitaatide vale kasutamine
argumenteerimata mõtteavaldus	argumenteerimata mõtteavaldus ja naiivsus
vasturääkivus	
liigitusvead	ühtse liigitusaluse puudumine lause grammatiliste vahendite ebaühtlus
väljendusvahendite üksluisus	väljenduse põhiliikide (jutustus, kirjeldus, dialoog) kasutamise ühekülgsus lauseehituse ja kõnekonstruktsioonide monotoonsus dialoogi primitiivsus kirjelduste primitiivsus ebaõige lauserõhk
üldsõnalisus, isikupära puudumine	keelestambid konkreetsuse puudumine ja fraasitsemine
stiilipuhtusvead	stiililine ebaühtlus, žargoon kirjanduslike eeskujude primitiivne kasutamine
kujundilisusvead	kujundite põhjendamatu rohkus vajaliku kujundilisuse puudumine ebatäpsed kujundid rahvapärase fraseoloogia väärtarvitamine
grammatilis-sõnastuslikud vead	sõnade vale järjekord aegade läbisegi tarvitamine otsese ja kaudse kõne segiajamine isikulise ja umbisikulise tegumoe läbisegi tarvitamine ühildumis- ja rektsioonivead.