doi:10.5128/ERYa13.13

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 13, 205–221

EESTI LASTEKEELE KORPUSE MORFOLOOGILISE MÄRGENDAMISE KITSASKOHTADEST

Kristiina Vaik, Virve-Anneli Vihman

Ülevaade. Artikli eesmärk on anda ülevaade sellest, mis raskendab avalikult kättesaadava eesti lastekeele korpuse automaatset morfoloogilist märgendamist ning anda soovitusi, kuidas tulevaste korpuste märgendamist ja standardiseerimist analüüsi tarbeks paremaks muuta. Analüüsisime korpust kirjakeelele mõeldud morfoloogiaanalüsaatori abil. Automaatse analüüsi järel vaatasime, kui suur osa sõnadest sai analüüsi või jäi analüsaatorile tundmatuks nii alamkorpuste kui lapse- ja hoidjakeele lõikes. Selgus, et analüüsi saanud sõnade osakaal igas alamkorpuses varieerus hoidjakeeles 94–98% ja lastekeeles 57–96% vahel. Suurt rolli mängib lindistuste üleskirjutamisviis: hoidjakeelt kirjutatakse üles kirjakeelele sarnaselt, kuid lastekeeles lähtutakse kuuldeortograafiast. Kõik alamkorpused küll järgivad CHILDES-i transkriptsioonisüsteemi ettekirjutusi, ent iga alamkorpus on koostatud erinevaid eesmärke silmas pidades ja on erineva transkribeerimisstiiliga, millest järjepidevalt kinni ei peeta. Märgenduse tulemust hindasime käsitsi läbivaatamise teel. Artiklis toome välja, millised olid nii tundmatuks jäänud kui vale analüüsi saanud sõnade sagedasemad probleemid ja pakume võimalikke lahendusi.

Võtmesõnad: lastekeel, korpus, automaatne märgendamine, transkriptsioon, eesti keel

1. Sissejuhatus

Suuremate korpuste loomise võimalikkus ning andmete kogumise, transkribeerimise ja märgendamise lihtsustumine on olnud tänapäeva keeleteaduse üks olulisemaid muutusi. Korpusanalüüs on aidanud kaasa ka lastekeele uurimise kvaliteedi tõusule. Esimesed lastekeele uuringud põhinesid päevikumärkmetel, kuid tehnoloogia areng andis aluse suurte spontaanse kõne andmekogude tekkimisele. Lapse spontaanse kõne salvestamine ja transkribeerimine võimaldab süstemaatiliselt

dokumenteerida ja analüüsida nii lapse kui lapsele suunatud kõnes esinevaid keelelisi nähtusi ja eripärasid. Andmekogude laiem kättesaadavus on tinginud vajaduse ühtse ja automaatse transkribeerimissüsteemi järele.

Korpuste puhul on oluline, et transkribeerimise ja märgendamise tasemel tehtavad otsused oleksid süstemaatilised, kuna need mõjutavad edasist informatsiooni kättesaamist (Behrens 2008: xx–xxii). Brian MacWhinney ja Catherine Snow lõid 1984. aastal arvutipõhise andmebaasi CHILDES, mis võimaldab keeleuurijatel oma keeleandmeid standardsel viisil transkribeerida, töödelda ja jagada (MacWhinney, Snow 1985). Lindistuste transkribeerimiseks ja märgendamiseks kasutatakse CHAT transkriptsioonisüsteemi, keelematerjali analüüsimiseks on võimalus kasutada CLAN tarkvara. Tänapäevaks on CHILDES muutunud lastekeele uurijate keskseks töövahendiks.

Võrreldes mõne teise keelega (nt soome, läti ja leedu keel pole CHILDES-is esindatud) on eesti lastekeele korpus mahu poolest heas seisus, kuid kvantitatiivset analüüsi ning võrdlemist hõlbustaks see, kui alamkorpused oleksid morfoloogiliselt märgendatud. Eesti keele analüüsimiseks on tehnoloogilised ressursid olemas ning tehtud on palju transkribeerimistööd, kuid automaatset märgendamist takistab lastekeele aga ka transkribeerimise ebastandardsus. Käesoleva artikli eesmärgiks on anda ülevaade sellest, mis raskendab automaatset märgendamist, hinnata ühe morfoloogilise märgendamise katse tulemusi ning anda soovitusi, kuidas tulevastes korpustes lihtsustada nende märgendatavust.

2. Eesti lastekeele korpus: 7 alamkorpust

Eesti laste suulise kõne salvestused on CHILDES-i andmebaasis olnud alates 1998. aastast. Seitse alamkorpust on oma nimed saanud korpuse koostajate järgi: Argus, Beek, Kapanen, Kohler, Kõrgesaar, Vija ja Zupping (CHILDES 2016). Ükski alamkorpus ei ole morfoloogiliselt analüüsitud. Alamkorpused erinevad üksteisest mitmes dimensioonis: andmete kogumise viis, korpuse eesmärk, laste arv ja vanus, korpuse suurus (vt tabel 1) ning hoidja ja lapse sõnade vaheline osakaal (vt joonis 1).[1]

Tabelist 1 on näha, et Kohler ja Kõrgesaar on läbilõikekorpused, kus andmed pärinevad salvestustest mitme lapsega. Enamikku alamkorpuseid iseloomustab pikiuurimusele omane lähenemine, kus teatud perioodi jooksul salvestatakse lapse ja hoidja(te) vestlusi. Vija on nn tihe korpus, kuna last lindistati 2- ja 3-aastaselt 6 nädala jooksul 6 tundi nädalas, vahepeal lindistati igakuiselt. Tiheda andmestikuga korpus annab ülevaatlikuma pildi lapse tegelikust keeleoskusest ning võimaldab suurema sõnavara salvestamist, lapse produktiivsuse uurimist ja teeb tõenäolisemaks, et keeles vähem sagedased nähtused on ka salvestatud materjalis esindatud (Tomasello, Stahl 2004).

Salvestatud vestlused toimuvad enamasti kodustes tingimustes mängimise või igapäeva toimingute ajal, kuid salvestusi on tehtud ka väljaspool kodu (nt lasteaias, muu pereliikme juures). Beeki korpus on loodud eelkõige lapsele suunatud keele (edaspidi hoidjakeele) uurimise eesmärgil.[2] Kõrgesaare alamkorpuses on andmete kogumiseks kasutatud nii piki- kui läbilõikeuurimuse põhimõtteid ning on loodud

[1] Artiklis esitatud analüüs on tehtud 2016. aasta kevade seisuga.

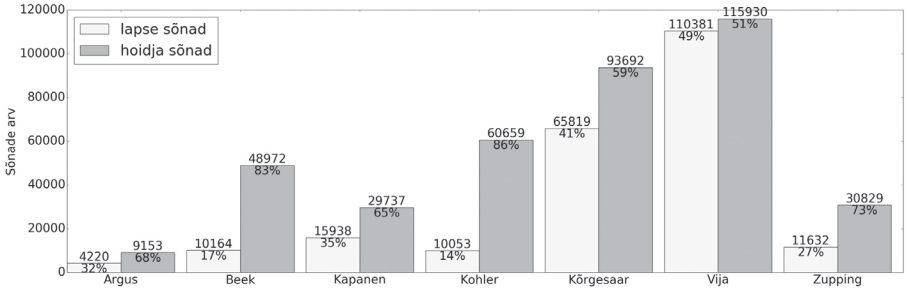
[2] Beeki alamkorpuse info eest täname retsensenti.

hoidjakeele uurimise eesmärgil. Kõrgesaare alamkorpus on ainus, kus on esindatud ka vanemad lapsed (vt tabel 1).

Tabel 1. Alamkorpuste info (CHILDES 2016)

Alamkorpus	**Laste arv**	**Vanus** (aasta;kuu)	**Lindistuste arv** (ja maht minutites, kui saadaval)
Argus	1	1;8–2;5	17 (185 min)
Beek	1	0;9–2;5	20
Kapanen	1	1;3–2;7	11 (470 min)
Kohler	8	0;11–2;3	61
Kõrgesaar	12	1;3–14;1	46
Zupping	1	1;3–4;2	23
Vija	1	1;7–3;1	74 (4335 min)

Korpuse koostamise üks põhimõte on, et see peaks olema võimalikult representatiivne ehk uuritava nähtuse suhtes esinduslik. Kui uurime lapse keelt, siis peame teadma, millist keelt ta kuuleb, seega on hoidjakeel oluline analüüsiobjekt lisaks lapse enda keelele (vt Argus 2010, Argus, Kõrgesaar 2014, Vihman 2015). Nii lapse kui hoidja keelekasutust uurides tuleb selgeks teha, kelle keelekasutust alamkorpus esindab. Joonisel 1 on kujutatud, kuidas jaotuvad hoidja- ja lastekeele sõnad[3] igas alamkorpuses. Hoidjakeel hõlmab endas kogu lapse ümbruses kasutatavat keelt (nt ema, isa, lindistaja jt), lastekeel hõlmab vaid ühe lindistatava lapse keelekasutust.



Joonis 1. Hoidja ja lapse sõnade jaotumine alamkorpustes (% on arvutatud iga alamkorpuse kohta eraldi)

Kõikides alamkorpustes on sõnade arvu poolest ootuspäraselt ülekaalus hoidja sõnad. Vija ja Kõrgesaare[4] korpuses jagunevad hoidja ja lapse sõnad enam-vähem võrdselt. Kõige suurem hoidja ja lapse sõnade jagunemise erinevus on Kohleri ja Beeki korpuses, kus domineerib hoidjakeel. Hoidja ja lapse sõnade jaotumise suured erinevused tulenevad laste vanusest (Beeki ja Kohleri korpuses on lindistatud kõige väiksemaid lapsi, vt tabel 1) kui ka lindistamise keskkonnast ning sessioonide arvust. Kõigis alamkorpustes on suurem osa transkriptsioonidest tehtud lastega vanuses 1 kuni 3 eluaastat.

[3] Kasutame terminit *sõna* tähenduses 'tühikute vahele jääv tähtede järjend' (st *laud*, *laua* on kaks erinevat sõna).
[4] Kõrgesaare korpusest on välja jäetud transkriptsioonid, mille osalejateks olid vaid täiskasvanud.

3. Alamkorpuste standardiseerimise probleemid

Selleks, et kvantitatiivselt uurida lapse morfoloogia ja süntaksi arengut ning sisendkeele (ehk hoidjakeele) mõju, on vaja märgendatud korpust. Praegune eesti lastekeele korpus võimaldab morfoloogiat uurida kindla lekseemi või lõpu abil, kuid mitte grammatiliste tunnuste abil. Näiteks ei ole võimalik otsida, kui palju kasutab laps minevikku või partitiivi. Morfoloogiliselt märgendatud kujul korpus võimaldaks teha otsinguid suurema haardega, nt otsida mõne morfoloogilise või süntaktilise näitaja põhjal või koondada lapse tehtud keelelisi vigu. Korpuste automaatne märgendamine ja analüüsimine ei ole triviaalne ülesanne ning eeldab, et juba transkribeerimise tasandil järgitaks teatud põhimõtteid. Praegune eesti lastekeele korpus on nende põhimõtete osas ebajärjekindel.

Korpus hõlmab endas (tüüpiliselt) kolme tüüpi informatsiooni: metaandmed, lingvistiline märgendus (ingl *annotation*) ja teksti elementide märgendus (ingl *corpus markup*) (McEnery, Hardie 2011: 30, vt ka Burnard 2014). Metaandmed sisaldavad informatsiooni teksti enda kohta – nt autor, aeg, keel, osalejad, vanus, sugu jpm. Lingvistilise märgendamisega pannakse paika teksti hierarhia (pealkirjad, lõigud, laused jm) ning vastavalt eesmärgile lisatakse kas käsitsi, pool- või täisautomaatselt märgendustasemed (nt morfoloogia, süntaks jm) (Muischnek jt 2003: 12–14). Teksti elementide märgendus kodeerib tekstisisest informatsiooni (nt millal kõneleja kõnevoor algab ja lõpeb). Oluline on selgeks teha, mida ja kuidas märgendada nii, et säilitada teksti algandmete kohta võimalikult palju infot ja et see oleks inim- ja masinloetav. (McEnery, Hardie 2011: 29–30)

Hennoste (2000: 92–93) kirjutab suulise kõne transkriptsiooni koostamise kahest printsiibist. Esimene on autentsus ehk transkriptsioonis peab säilima informatsioon, mis on suhtluse loomuse suhtes tõene. Teine printsiip on praktilisus ehk transkribeerimise tavad peavad olema andmete korraldamise ja analüüsi viisi suhtes kasulikud. Nendest printsiipidest lähtuvalt tuleb märgendada neid nähtuseid, mida uurijal on tarvis, kuid et see üldist pilti üle ei küllastaks. Selles artiklis käsitleme vaid morfoloogilist märgendamist, mis võiks hõlbustada keeleteadlaste üldiselt kasutatavaid otsinguid, kuid jätame kõrvale muud võimalikud sõltread, nt infostruktuur, argumentstruktuur, lapse tähelepanu ja liigutuste informatsioon jpm.

CHILDES-i süsteemi järgi algavad transkriptsioonid päisega, kus antakse informatsiooni lindistuse aja, koha, osalejate, kestuse, laste vanuse jms kohta (vt näide 1).

(1) @UTF8
@PID: 11312/c-00026785-1
@Begin
@Languages: est
@Participants: CHI Andreas Target_Child, MOT Maigi Mother
@ID: est|Vija|CHI|3;0.||||Target_Child|||
@ID: est|Vija|MOT|||||Mother|||
@Date: 26-FEB-2001
@Situation: päeval elutoas, mängimine ja magnettahvlile joonistamine (Vija/30000, vanus 3;0)[5]

[5] Näidete alla on lisatud korpuse nimi, transkriptsioonifail ja vaatlusaluse lapse vanus.

Põhiridadele paigutatakse kõnelejat tähistav kolmetäheline kood, millele järgneb kõne ning sellele lisatakse põhi- või sõltreale juurde, kas transkribeerija- või uurijapoolsed kommentaarid või märgendused (Argus 2007: 68, vt näiteid 2, 3 ja 4). Näide (4) on Browni korpusest, kus morfoloogilist ja süntaktilist infot on esitatud eraldi sõltridadel.

(2) *MOT: arvuta need kõigepealt ära.
*CHI: jah mm kaheksa miinus seitse on üks.
*CHI: niimoodi kümme miinus üks on üheksa.
%com CHI kirjutab ja ise räägib samal ajal kaasa.
(Korgesaar/Gregory/gregory03, vanus 7;8)

(3) *FAT: mida sa tahad kätte , issi ei tea , kus see teritaja on .
*MOT: see teritas väga ilusasti muidu .
*CHI: telita [*] .
%err: terita=teritaja $MOR
%par: CHI aevastab
(Vija/20008, vanus 2;0)

(4) *CHI: big drum .
%mor: adj|big n|drum .
%gra: 1|2|MOD 2|0|INCROOT 3|2|PUNCT
*MOT: big drum ?
%mor: adj|big n|drum ?
%gra: 1|2|MOD 2|0|INCROOT 3|2|PUNCT
(Brown/Adam/adam01, vanus 2;3)

Lindistuste transkribeerimiseks kasutatakse kuuldeortograafiat, mis ei anna küll tõetruud pilti sellest, milline on lapse tegelik keelekasutus, kuid transkriptsiooni loetavuse mõttes on see eesti keeles ainuõige lähenemine. Keelematerjali analüüsimiseks saab kasutada CLAN tarkvara, millega on näiteks võimalik otsida sõna ja sõnaosa sagedust ja esinemise kontekste (otsitud sõna lausungit ja ka sellele eelnevaid/järgnevaid lausungeid), keskmist lausungi pikkust jm, kuid puudub automaatse morfoloogilise analüüsi võimalus, sest CLAN-i morfoloogiaanalüsaator ei ole eesti keele tarbeks rakendatav.

Suulise kõne automaatne analüüsimine on keeruline, kuid lapse suulise keele analüüsimine on veelgi keerulisem (vt Behrens 2012, Aviad jt 2013). Lindistatud on spontaanset suulist kõnet, mis sisaldab elemente, mida pole tarvis analüüsida, nt häälitsused. Selleks, et korpuseid oleks võimalik analüüsida nii, et need annaksid keelekasutuse kohta autentse pildi, ja et oleks võimalik neid standardsele kujule viia, tuleb alustada juba korpuse tekstide transkribeerimise tasandist (Argus 2007: 71). Eesti lastekeele korpuste koostajad tunnistavad vajadust ühtsete transkribeerimise põhimõtete järele, sest vaid nii on analüüsitulemused usaldusväärsed ja omavahel võrreldavad.[6] Korpuste koostajad on suurimate murekohtadena välja toonud kõnevooru pikkuse kindlakstegemise, eneseparanduste ja -täienduste lahendamise, märgendite kasutamise (millised ja kui palju) ning venituse, kokkuhäälduse, intonatsiooni ja emotsiooniga öeldu ülesmärkimise (Kask 2016: 20).

[6] Andmed pärinevad Paula Helena Kase (2016) bakalaureusetöö raames läbi viidud küsitluse kokkuvõttest, mis põhines Reili Arguse, Maigi Vija, Sirli Zuppingu ja Helen Kõrgesaare kirjalikel vastustel.

Lapse puhul on tegemist areneva keelekasutusega, milles esineb palju erilisi tunnuseid. Näiteks kui kordustest koosnevas lausungis eraldatakse sõnu komade abil (**CHI: onu , onu , onu*), siis sellise üleskirjutusviisi järgi koosneb lapse lausung kolmest erinevast sõnast. Kui aga transkribeerida seda lausungit [/] abil (**CHI: onu* [/] *onu* [/] *onu* [/]), siis koosneb see ühest sõnast, sest CLAN kohtleb seda kui korduvat üksust.

4. Vead ja nende tähistamine

Lastekeeleuurijad huvituvad vigastest vormidest, sest need heidavad valgust lapse keelelisele arengule ja teadmistele. Sellest lähtuvalt toimub ka vigaste vormide transkribeerimine ja märgendamine. CHILDES-i transkriptsioonisüsteemis märgendatakse vigu kolmel erineval viisil: [: *sihitud_sõna*], [= *tähendus/selgitus*] ja [*]. Esimesel juhul lisatakse nurksulgudesse kooloni järele sõna kirjakeelne vorm. Teisel juhul lisatakse nurksulgudesse võrdusmärgi järele moodustatud sõna korrektne vorm või tähendus. Näites (5) on ühes lausungis kaks häälduslviga tähistatud erineval viisil.

(5) *FAT: kriit pane tahvli peale .
*CHI: kit [: kriit] .
*CHI: kit [: kriit] (.) vahvlile [= tahvlile] pääle [: peale] .
(Vija; 20007, vanus 2;0)

Kooloni kasutamise eelis vea tähistamisel seisneb selles, et öeldud sõna asendatakse selle kirjakeelse vormiga, kuid seejuures jääb püsima nende semantiline samaväärsus. Võrdusmärgi kasutamist [= *tähendus/selgitus*] soovitatakse CHAT-reeglistikus siis, kui põhireal soovitakse öeldule anda lühikest seletust või tähendust. Näites (5) on tegemist fonoloogilist laadi vigadega: hääliku ärajätt *kit* 'kriit' ja asendus *vahvlile* 'tahvlile'. Näiteks kui otsida sõna *kriit*, siis tagastatakse ka lapse öeldud *kit*. Aga kui otsida sõna *tahvlile*, siis võrdusmärgiga tähistamisel ei seostata seda sõnaga *vahvlile*. Seega, otsing ei leia sellises veamärgenduses üles kõiki selle sõna esinemisi ja häälduslvariante. Näites (6) kasutatakse võrdusmärki mitte vea tähistamiseks, vaid selleks, et täpsustada/selgitada öeldut, *opsas* tähenduses 'süles'.

(6) *MOT: jah ole minu opsas [= süles] ole minu süles
(Zupping; 1_03.03, vanus 1;3)

Kolmanda vea tähistamise viisina kasutatakse põhireal sümbolit [*] ja sõltreale võib lisada vearea (*%err*), kus vigase vormi järel on korrektne vorm. Näites (7) on tegemist vale käände valikuga ning $MOR tähistuskoodi abil on selliseid morfoloogilisi vigu võimalik CLAN-is otsida, aga vaid siis, kui need vead on juba transkriptsioonis eelnevalt märgendatud. Sümboli [*] abil on vigu tähistatud vaid Vija alamkorpuses (vt tabel 2).

(7) *CHI: issi , loe seda .
*CHI: issi , nüüd see [*] ei pane kinni !
%err: see=seda $MOR
(Vija; 20007, vanus 2;0)

Nagu juba eeltoodud näidetes näha, pole eesti lastekeele alamkorpuste koostajad olnud veatähistamise osas järjekindlad: kord on viga märgendatud ühtmoodi, kord teistmoodi ja vahel üldse mitte, vt näide (8) (vt ka Argus 2008).

(8) *FAT: viskad minema või?
*FAT: kus sa viskad selle?
*CHI: kinn.
*FAT: sinna viskad jah.
(Kõrgesaar; arabella01f, vanus 1;8)

Tabel 2 annab ülevaate sellest, kui palju ja missuguseid veamärgendusi on alamkorpustes kasutatud. Hõlmatud on ainult need vead, mis on ühel või teisel viisil fikseeritud veaks.

Tabel 2. Vea märgendamine alamkorpustes

Alamkorpus	**[*: sihitud_sõna*]**	**[= *tähendus/selgitus*]**	**[*]**
Argus	522	–	–
Beek	43	–	–
Kapanen	9	1198	–
Kõrgesaar	280	1219	–
Kohler	2062	173	–
Vija	8534	3122	547
Zupping	–	2883	–

Vigade märgendamisel on kaks suuremat probleemi. Esimene seisneb selles, et vigu ei märgendata ühtselt, st transkriptsioonis võib üks ja seesama viga olla tähistatud kord ühel, kord teisel viisil. Teine probleem on keerulisem: vigade märgendamisel ei olda järjepidevad, st üks ja seesama viga võib kord olla tähistatud, kord mitte. Kui viga on jäetud märgendamata (või kui seda on tehtud mitmel erineval viisil), siis seda ei ole võimalik automaatselt tuvastada. Näiteks kui otsida sõna *sinna*, siis ei leita lapse vormi *kinn* (vt näide 8). See on suurem probleem, sest meil puudub tegelikkuses info, kui palju vigaseid vorme on jäänud kogu korpuses tähistamata.

5. Morfoloogilise märgenduse hindamine

Eelnevalt kirjeldatud probleemidest lähtuvalt püstitasime eesmärgi hinnata korpuste märgendatavust ja kaardistada mõned olulisemad kitsaskohad. Selleks tegime esmase morfoloogilise märgendamise katse ja analüüsisime selle tulemusi (lähemalt võib lugeda Vaik 2016).

5.1. Meetod

Algandmetena kasutasime CHILDES-i poolt eelnevalt automaatselt konverteeritud eesti lastekeele korpuse XML-faile ning lisasime neile morfoloogilise tasandi. Morfoloogilise analüüsi käigus lisasime iga sõna kohta infot selle lemma, sõnaliigi ja morfoloogiliste kategooriate kohta: käändsõnal arv ja kääne, tegusõnal pööre,

tegumood, aeg, kõneviis, kõneliik. Teksti morfoloogiliseks analüüsimiseks kasutasime eesti keele morfoloogiaanalüsaatorit *etana*. Morfoloogiline analüüs koosneb kahest etapist – üksiksõnade analüüsimine ja nende ühestamine (Kaalep, Vaino 2000: 89).

Analüüsi kvaliteeti on võimalik tõsta analüsaatori allkeelespetsiifilisemaks muutmise teel, nt kasutada kasutajasõnastikku (vt Muischnek jt 2011: 113). Sel juhul kontrollib analüsaator, kas sõna esineb kasutajasõnastikus või mitte. Esinemise korral võetakse selle analüüs sealt, puudumise korral tehakse morfoloogiline analüüs. Kasutajasõnastikku saab panna sõnu, mida analüsaator muidu analüüsida ei suudaks, ja sõnu, mis peaksid konkreetses tekstis teistsuguse analüüsi saama.

Lastekeele korpust analüüsisime ilma analüsaatorit kohandamata, lisaks ei teostatud oletamist ega ühestamist, st tundmatute sõnade korral tagastatakse #### ja õnnestunud analüüsi korral jäetakse alles kõik analüüsivariandid. Meeles peab pidama seda, et analüsaator on loodud eesti kirjakeele tarbeks, kuid lapse ja hoidja keelekasutus erineb normeeritud kirjakeelest leksikaalsete ja ortograafiliste eripärade poolest. Seega oleks kasutajasõnastiku kasutuselevõtt lastekeele korpuse analüüsimisel kindlasti kohane. See oli meile teadaolevalt esimene katsetus olemasolevate vahenditega morfoloogiliselt analüüsida eesti lastekeele korpust.

5.2. Automaatse analüüsi tulemused

Nii laste- kui hoidjakeelele on iseloomulik mitteformaalne kõne ja emotsionaalselt lähedane kõnelemise situatsioon, kuid neid tuleks oma erinevate tunnuste tõttu eraldi vaadelda. Võtame esmalt vaatluse alla hoidja sõnad. Analüüsi saanud kui ka analüsaatorile tundmatuks jäänud sõnade osakaal igas alamkorpuses on üsna stabiilselt jaotunud. Analüüsi saanud sõnade üldine osakaal varieerub alamkorpustes 92–98% vahel ja tundmatute sõnade osakaal 2–8% vahel. Tabelis 3 on näide, kuidas jaotuvad Kohleri korpuses analüüsi saanud ja analüsaatorile tundmatuks jäänud hoidja ja lapse sõnad.

Tabel 3. Kohleri korpuse sõnade analüüsi ülevaade

Lapse vanus kuudes	Lapse sõnad (arv ja %)			Hoidja sõnad (arv ja %)		
	analüüsitud	tundmatud	kokku	analüüsitud	tundmatud	kokku
0–11	6 (100%)	0 (0%)	6 (100%)	281 (95%)	14 (5%)	295 (100%)
12–23	4592 (93%)	348 (7%)	4940 (100%)	40 724 (97%)	1160 (3%)	41 884 (100%)
24–35	4992 (98%)	115 (2%)	5107 (100%)	18 224 (99%)	256 (1%)	18 480 (100%)
Kokku	9590 (95%)	463 (5%)	10 053 (100%)	59 229 (98%)	1430 (2%)	60 659 (100%)

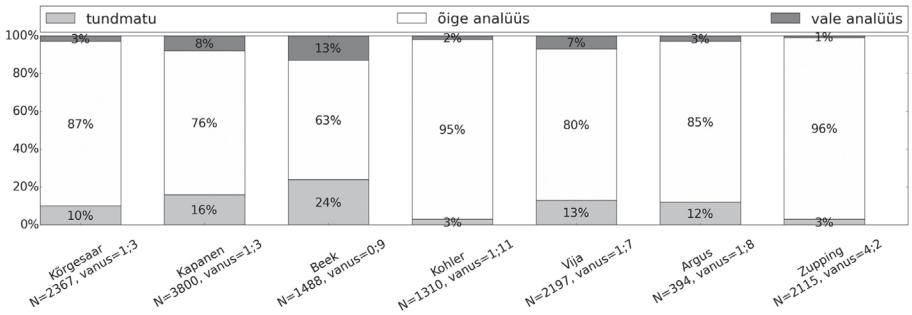
Need tulemused on lootustäratavad, kuna tegemist on suulise keelega, mille erijooned võivad olla kirjakeele analüüsimiseks loodud morfoloogilise analüsaatori jaoks problemaatilised. Näiteks uue meedia keelekasutus (ehk internetikeel) on

oma spontaansuse ja mitteformaalsuse tõttu sarnane suulisele keelele ja erineb kirjakeelest nii leksikoni kui ortograafia poolest. Uue meedia alamkorpuste esmasel morfoloogilisel analüüsimisel saadi 6–27% tundmatuid sõnu (Muischnek jt 2011). Pärast kasutajasõnastiku ja eeltöötluse rakendamist vähenes tundmatute sõnade osakaal 5–11%-ni. Seega, väike tundmatute sõnade % hoidjakeeles on hea ning need andmed viitavad sellele, et korpuse transkribeerijad on hoidjakeelt üles märkinud kirjakeelele sarnaselt.

Laste puhul sõltuvad analüüsi saanud ja analüsaatorile tundmatuks jäänud sõnad (ja ka sõnade koguarv) eelkõige lapse vanusest. Analüüsitud sõnade üldine osakaal varieerub alamkorpustes 57–96%. Kõige suurem analüsaatorile tundmatuks jäänud sõnade osakaal on Beeki alamkorpuses (34–74%, varieerumine vanusega) ja kõige väiksem tundmatute sõnade osakaal Vija (3–22%) ja Kohleri (0–7%) alamkorpuses. Olenemata sellest, et alamkorpustes on tundmatute sõnade % amplituud suur, jääb üldine skoor kogu korpuses alla 10%.

Tulemuste tõlgendamisel ning lapse ja hoidja keelekasutuse kvantitatiivsel hindamisel tuleb arvestada korpuse andmete kogumisviisi ja suurusega. Väikse andmestikuga ja erinevate andmete kogumisviisidega korpuseid ei ole võimalik omavahel võrrelda, sest see viib valede järeldusteni. Võib väita, et kogu korpuses üldjuhul kahaneb vanuse suurenemisega tundmatute sõnade osakaal, kuid leidub ka erandeid. Beeki alamkorpuses on 0–11-kuuse lapse tundmatuks jäänud sõnade % väiksem kui 1-aastase seas (69% *vs.* 74%).

Seni oleme vaadelnud vaid analüsaatorile tundmatuks jäänud sõnu, kuid analüüsi tulemuste adekvaatsuse hindamiseks tuleb vaatluse alla võtta ka analüüsi saanud sõnad. Analüüsitud sõnade puhul peab uurima, kas morfoloogiline analüüs on täpne. Käsitsi läbivaatamiseks valisime juhuslikkuse alusel igast alamkorpusest ühe transkriptsiooni ja hindasime iga sõna puhul, kas saadud analüüs on õige või mitte. Joonisel 2 on kujutatud, kuidas jaotuvad alamkorpuste transkriptsioonides tundmatuks jäänud, õige ja vale analüüsi saanud nii lapse kui hoidja sõnad. Kõige paremad tulemused olid ootuspäraselt Zuppingu transkriptsioonis, kuna keelematerjal pärineb lindistusest lapsega vanuses 4;2. Kõige enam oli vale analüüsi saanud sõnu Beeki transkriptsioonis (laps vanuses 0;9). Üllatuslikult head tulemused olid Kõrgesaare (laps vanuses 1;3) ja Kohleri (laps vanuses 1;11) transkriptsioonides.



Joonis 2. Tundmatud, õige ja vale analüüsi saanud sõnad alamkorpuste juhuvalimi transkriptsioonide põhjal

Need arvud sunnivad küsima: millest tekivad alamkorpuste vahel nii suured erinevused? Osalt on põhjus väikses sõnade koguarvus ehk korpuse suuruses. Teisalt viitavad need andmed sellele, et analüsaatorile tundmatuks jäänud sõnade osakaal sõltub üleskirjutamise viisist. Transkribeerija peab märgendused lisama transkriptsiooni nii, et vead oleksid juba esimesel tasandil liigitatud (Argus 2007: 74). Teisisõnu, transkriptsioonides on oluline, et oleks tähistatud öeldud vorm kui ka vorm, mida üritati öelda.[7] Nii on võimalik analüüsida lastekeelt, säilitades selle omapära ja normist kõrvalekaldumine. Selliste eksimuste uurimine on oluline, sest need pole (alati) juhuslikud, vaid peegeldavad lapse arusaamu vormimoodustussüsteemi mustritest ja mallidest. Vaatame taas Kohleri alamkorpuse salvestust (laps vanuses 0;11), kus pole ühtki tundmatuks jäänud sõna (vt tabel 4).

Tabel 4. Lapse sõnad Kohleri alamkorpuses (vanus 0;11)

Lapse vorm	**Asendus [: *sihitud_sõna*] abil**	**Analüüs**
teh	tere	_l_ tere+0
tehe	tere	_l_ tere+0
täh	aitäh	_l_ aitäh+0
äh	aitäh	_l_ aitäh+0

Lapsel on sõnu kokku 6. Kontekstita need vormid justkui ei tähenda midagi ning jääksid analüsaatorile (ühtlasi ka lugejale) tundmatuks. Kuna nendele vormidele on juurde lisatud märgend [: *sihitud_sõna*] koos kontekstist tuleneva sihitud sõnaga, mida laps üritas öelda, siis nii saab analüsaator oma tööga hästi hakkama ja seetõttu pole selle lapse keelekasutuses jäänud ükski sõna analüsaatorile tundmatuks.

Kapaneni ja Beeki alamkorpuses on tundmatute ja vale analüüsi saanud sõnade osakaal kõige suurem ja just nendes korpustes kasutatakse kõige enam nii lapse kui hoidja kõne transkribeerimisel kuuldeortograafiat (kuid mitte järjepidevalt). Kapaneni ja Beeki alamkorpuses kasutatakse läbisegi *head isu* ja *ead isu*, *präegu* ja *praegu*, *jaaah* ja *jah*, *äitäh* ja *aitäh* jm. Sellise kirjapildiga antakse edasi lapse hääldust, kuid valesti produtseeritud sõnad muutuvad kättesaamatuks nii CLAN-i otsingutele kui ka analüsaatorile. Seetõttu on otstarbekam transkribeerida lapse häälduspärast vormi, kuid selle kõrval märgendada viga, kasutades asendust [: *sihitud_sõna*]. Ainult häälduspärase vormi üleskirjutamine mõjutab analüsaatori väljundit nii tundmatute kui ka vale analüüsi saanud sõnade osas, nt sõna *ead* (*head*) saab analüüsiks *iga+d _S_ Pl Nom* (vrd *hea+d _A_ Pl Nom, Sg Par* või *hea+d _S_ Pl Nom, Sg Par*), sõna *präägust* (*praegust*) saab analüüsiks *prääk+o _S_ Sg Gen* (vrd *praegune+t _A_ Sg Par*).

Järgnevas vaatame valesti analüüsitud sõnu lähemalt.

6. Valesti analüüsitud sõnad

Täielikult vigadeta morfoloogiliselt märgendatud korpus eeldab, et iga sõnavorm saab õige sõnaliigilise kuuluvuse, käändsõnad õige arvu ja käände, verbid õige arvu, isiku, tegumoe, aja, kõneviisi ja kõnelaadi. Transkriptsioonide käsitsi läbivaatamise käigus hindasime, kas tegu on õige lemma, sõnaliigi ja morfoloogiliste

[7] Selle hindamine mõnikord sõltub transkribeerija intuitsioonist ja tõlgendusvõimest, sest lapse teadmised keelest on ebatäiuslikud, mistõttu ta alati ei püüagi produtseerida seda "õiget" vormi.

kategooriatega. Vale analüüsi saanud sõnade puhul oli sageli raske nende sõnaliigilist kuuluvust määrata, sest tihtipeale polnud isegi konteksti olemasolul aru saada, millega tegu. Kui see valmistab juba inimesele probleeme, siis on sellega probleeme ka analüsaatoril, sest tegu on kirjakeelest hälbiva keelekasutusega. Vale analüüsi saanud sõnad paigutasime 5 erinevasse rühma: onomatopoeetilised sõnad, häälitsused, pärisnimed, vale lemma või sõnaliik ning ebastandardse ortograafiaga sõnad.

6.1. Onomatopoeetilised sõnad

Kõik helijäljenduslikud sõnad on nn kirvemeetodil ühte rühma paigutatud, sest nende sõnaliigilist kuuluvust on raske määrata, nt *kiiga*, *kõps*, *nämm*, *patsu*, *sulla*, *summ*, *tiks*, *viuviu*.

Onomatopoeetiliste sõnade rolli on alatähtsustatud, kuid olenemata sellest, kas keel on häälikusümboolika poolest rikas või mitte, kuuluvad onomatopoeetilised sõnad lapse esimeste sõnade hulka ja on ka hoidjakeeles sagedased (Laing 2014a, 2014b). Reili Argus eristab onomatopoeetiliste sõnade hulgas ka *imitatiive*: onomatopoeetilised sõnad, mille häälikuline kuju võib olla varieeruv, kuid ei muutu morfoloogiliselt, nt kiirabiauto signaali imiteeriv *viuviu*, kõndimise väljendamiseks kasutatav *tipa-tapa* (Argus 2004: 19–22).

Oluline küsimus on see, kuidas onomatopoeetilisi sõnu tuvastada. CHAT transkriptsioonisüsteem soovitab onomatopoeetiliste sõnade lõppu lisada sümbolid *@o –> kõps*@o, *nämm*@o (Argus 2007: 72–73; vt ka CHAT), aga eesti alamkorpustes kasutatakse seda tegelikkuses väga vähe. Sellise märgenduse kasutamine teeks onomatopoeetilised sõnad nähtavaks, mistõttu oleks võimalik neid ka automaatselt kasutajasõnastikku lisada. Ainsana on onomatopoeetilisi sõnu märgendatud Vija alamkorpuses (781 korda), kuid mitte järjekindlalt, st sama onomatopoeetiline sõna esineb nii märgendatud kui märgendamata kujul.

Onomatopoeetiliste sõnade sõnaliigilise mitmesuse tõttu tuleks mõelda kasutajasõnastikus eraldi sõnaliigi defineerimise peale. Eesti keele käsiraamat (EKK) jagab tähenduse järgi onomatopoeetilised sõnad interjektsioonide alla. Hennoste (2002: 67) nimetab jällegi interjektsiooni sõnaliigiliseks prügikastiks, kuhu on paigutatud üksused, mis mujale ei sobi. Tema arvates on onomatopoeetilised sõnad interjektsioonide alla paigutatud sellepärast, et neil on kaldeline foneetiline ja fonoloogiline struktuur ning nad paiknevad sõna ja mittesõna piirimail. Onomatopöa on sageli ka süntaktiliselt seostamata, st lauses eraldiseisev. Näites (9) ja (10) jääb imitatiivi sõnaliigiline kuuluvus segaseks:

(9) *CHI: addrr, drrr, brrr
*MOT: just, niimoodi sa õues sõidad vankriga
(Argus 2004: 27)

(10) %comment: osutab autole paberil
*MOT: nii, tuled teen
*MOT: sina tee katusele
*CHI: iiuiiu
%comment: Hendrik joonistab vilkureid
(Argus 2004: 28)

Argus analüüsib neid nimisõnaks või verbiks (Argus 2004: 28), kuid neid võib analüüsida isoleeritud onomatopoeetilisteks sõnadeks. Väidetakse, et ühesõnaliste lausungite perioodil ongi raske sõnu liigitada ja sõnaliikidest saab alles siis rääkida, kui laps hakkab kasutama mitmesõnalisi väljendeid. Kui lapse lausung on morfoloogiliste tunnusteta, siis pole analüüsimiseks ka laiemast kontekstist kasu.

6.2. Häälitsused

Need on järjendid, mille tähendusest pole võimalik aru saada ka konteksti olemasolul, nt *eo*, *kookai*, *manni*, *mm*, *muks*, *op*, *paa*, *s*, *t*, *ä*, *ämm*, *änn*, *öö* jm.

Sellised sõnad on analüsaatori jaoks problemaatilised. Nt *t*, *op*, *mm*, *ä*, *s* analüüsitakse lühenditeks. Selliste ühe- või mitmetäheliste "sõnade" valesti analüüsimise vältimiseks ja tuvastamiseks piisaks, kui kasutada spetsiifilist märgendust *@k* (mitme tähe jaoks) või *@l* (ühe tähe jaoks). *@l* märgendust on kasutatud Vija (343 korda), Zuppingu (3 korda) ja Kohleri (41 korda) alamkorpustes. Järjend *eo* saab analüüsiks *idu+0 _S_ Sg Gen*, kuid konteksti vaadates saab aru, et laps ei räägi idudest, vaid tegemist on silbitamisega. Ja selliste sõnade nagu *kaka*, *öö*, *ämm* puhul on tegemist häälitsustega, kuid analüsaatori jaoks näevad need välja kui üldnimisõnad ja seetõttu saavad vale analüüsi.

6.3. Pärisnimed

Pärisnimed on korpustes üldjuhul eristatavad suure algustähe abil, nt *Annika*, *Antsu*, *Carlos*, *Liisu*, *Sirts* jm. Pärisnimede valesti analüüsimise vastu ei saa üleskirjutaja midagi teha, sest need puuduvad analüsaatori kasutajasõnastikust, aga et need vormilt sarnanevad üldnimisõnadega, siis saavad need ka vale analüüsi. Pärisnimede üleskirjutamisel tuleks järgida CHAT formaadi ettekirjutusi ehk transkribeerimisel kasutada suurtähte, ning selle põhjal võib analüsaatori tarbeks lisada pärisnimed kasutajasõnastikku.

6.4. Vale lemma või sõnaliigiga sõnad

Vale analüüsi on saanud nt *mine-mine*, *musi-musi*, *mõmmi*, *nuku*, *siukse*, *tantsi-tantsi*, *venna*, *väga-väga* jm, samuti *kuule*, *palun*, *näe*. Viimased sõnad paiknevad verbi ja interjektsiooni piirimail ning oleksid justkui tekkinud täistähenduslike sõnade muutumise teel.

Selles rühmas on palju reduplikatiivseid sõnu, mille puhul on analüsaator õigesti analüüsinud nende sõnaliigi ja morfoloogilised kategooriad, kuid on andnud vale lemma, nt *väga-väga+0 _D_* 'väga-väga', *mine-mine+0 _V_ Pers Prs Imprt Sg2* 'mine-mine'. Lisaks on seal sellised astmevaheldusliku sõnad, mis on nihutatud astmevahelduseta tüüpi, nagu *mõmmi*, *venna*, *nuku*, mida analüüsitakse genitiivivormideks. Valeanalüüside vältimiseks on kaks võimalust. Esimene võimalus oleks kasutada asendust [: *sihitud_sõna*]. Teine võimalus on kohandada analüsaatorit ehk lisada neid sõnu kasutajasõnastikku ning nende analüüsimisel

rakendada ühestamist (st kõikvõimalikest interpretatsioonidest valitakse antud konteksti sobiv analüüs).

6.5. Ebastandardse ortograafiaga sõnad

Mõned sõnad võivad olla läbinud teatud häälikuteisendused, nt *keti* 'kõdi', *kispi* 'küpsis', *laua* 'laulma', *mammu* 'mari', *eita* 'ei taha', *tahta* 'tahan' jm. Vormid *kispi*, *mammu* ja *tahta* ei ole olemuselt samasugused vead, kuid põhjustavad analüsaatorile sarnast probleemi. *Mammu*[8] on vorm, mida kasutataksegi lapsega rääkides. *Kispi* on lapse püüd öelda õiget sõna. Vormi *tahta* kasutatakse vigaselt. Neid vorme iseloomustab see, et neile kõigile järgneb transkriptsioonides märgend [= *tähendus/selgitus*].

Selle rühma sõnade puhul peab tagasi pöörduma vigade ehk eksimuste üleskirjutamisviisi juurde. Vealiigitamisest on oluline rääkida sellepärast, et see mõjutab otseselt analüsaatori tööd. Veamärgend peaks vigasele vormile vahetult järgnema (nt *pitti* [: *pilti*], *mõh* [= *mõmmi*] *kodu*), kuid selle ettekirjutuse järgimine pole alamkorpustes olnud järjepidev, nt *õmmi teep* [= *mõmmi teeb*] (veamärgend peaks järgnema igale valele sõnale eraldi). Nendes sõnades on sageli kasutatud võrdusmärgiga veamärgendust, mistõttu saavad need ka vale analüüsi, kuna programm ei oska arvestada sedalaadi veatähistamisega.

Kui vaadata alamkorpustes tundmatuks jäänud sõnu ja üleskirjutaja vealiigitamist (vt tabel 2), siis võib märgata teatud seost. Korpustes, kus kasutatakse kooloniga tähistatud veamärgendust, on ka üldjuhul vähem tundmatuks jäänud sõnu, nt Vija, Kohleri ja Kõrgesaare alamkorpus. Zuppingu alamkorpuses ei esine kooloniga veamärgendust, mistõttu on ka lapse tundmatuks jäänud sõnade osakaal suur (vt Vaik 2016). Alamkorpuste praeguse kuju läbivaks jooneks on see, et laste keelekasutust märgitakse üles häälduspäraselt ja hoidjakeelt kirjakeelele sarnaselt.

Õige analüüsi valimine läheb keeruliseks siis, kui sõna paikneb kahe sõnaliigi vahel või kasutatakse teise sõnaliigi funktsioonis. Suur osa kategooriatest on vormi põhjal üheselt määratavad, kuid on selliseid mitteühesuse tüüpe, mis valmistavad isegi inimühestajale raskusi, nt käändsõnad ja verbid, mille vormidest arenevad adpositsioonid ja adverbid (nt *kätte*, *käes*, *alates*), verbi ja adjektiivi piirimail paiknevad partitsiibid (nt *surnud*, *kadunud*) ning adverbi ja konjunktsioonide piirimail paiknevad sõnad (nt *aga*, *nagu*, *kui*). Morfoloogilise analüüsi mõttes oleks hea, kui sellist piirimail asetsemist oleks võimalikult vähe ja seetõttu peaksid sõnaliigid olema kirjeldatud nii, et ka süntaksit saaks võimalikult otstarbekalt kirjeldada (vt Muischnek, Vider 2005: 102–104, Kaalep jt 2000: 627–631).

Õige analüüsi ja veamärgenduse valimisel peab olema ettevaatlik, kuna selle käigus võidakse omistada lapsele tähendusi ja mõisteid, mida ta tegelikkuses pole omandanud ega oska kasutada. Sama oht kehtib ka transkribeerimisel, kui lapse keelelisele väljendusele omistatakse teatud fonoloogiline või morfoloogiline kuju, mida laps pole produtseerinud ega omandanud. Transkribeerimine nõuab palju tähelepanu ja süvenemist, mistõttu see olekski õige etapp, mil järgida konkreetseid, ühtseid juhiseid, mis lubaksid ebastandardseid kohti paremini kategoriseerida ning märgendamise protsessi tõhusamaks muuta.

[8] Vormi *mammu* on võimalik markeerida märgendite *@c* (vorm, mida laps kasutab) ning *@f* (vorm, mida peres kasutatakse) abil.

7. Kokkuvõte

Uurimuse eesmärk oli hinnata eesti lastekeele korpuse morfoloogilise märgenduse katse kvaliteeti, kaardistada märgendamisega seotud olulisemaid probleeme ning anda soovitusi, kuidas korpuse transkribeerimist ja standardiseerimist morfoloogilise analüüsi tarbeks paremaks muuta. Korpuse automaatse morfoloogilise analüüsimise järel vaatasime, kui suur osa sõnadest sai analüüsi või jäi tundmatuks nii alamkorpuste kui lapse- ja hoidjakeele lõikes. Analüüsi käsitsi läbivaatamise tulemusena selgitasime välja, millised olid sagedasemad probleemid tundmatuks jäänud ja vale analüüsi saanud sõnadel.

Selgub, et hoidjakeele analüüsitud sõnade osakaal igas alamkorpuses varieerub 94–98% vahel. Arvestades seda, et tegemist on suulise keelega, on need tulemused väga head. Väike tundmatute sõnade % hoidjakeeles viitab sellele, et korpuse transkribeerijad on hoidjakeelt üles märkinud kirjakeelele sarnaselt.

Lastekeeles on analüüsi saanud sõnade varieerumine suurem, 57–96% vahel. Ka lastekeeles mängib suurt rolli üleskirjutamise viis: ülesmärkimisel kasutatakse palju kuuldeortograafiat, kuid seejuures jäetakse märkimata, mida laps tegelikult öelda tahtis. Vead on olulised tähised lapse keelelise arengu iseloomustamiseks ja nende ülesmärkimine on tähtis, sest vahel ei piisa ka kontekstist arusaamaks, mis sõnaga on tegu. Samas vea ülesmärkimisega võib kaasneda ka oht, et transkribeerija püüab lapse poolt öeldut ületõlgendada. Täielikult vigadeta morfoloogiliselt märgendatud korpus eeldab, et iga sõnavorm saab õige sõnaliigilise kuuluvuse ja morfoloogilise analüüsi, kuid lastekeeles on palju selliseid sõnu, mis pole vormi põhjal üheselt määratletavad. Käsitsi hindamise käigus paigutasime vale analüüsi saanud sõnad rühmadesse, mille hulgast kerkisid esile need, mis ei sobitu traditsiooniliste sõnaliikide kategooriate hulka (nt onomatopoeetilised sõnad ja häälitsused).

Morfoloogiaanalüsaatorit on võimalik kasutajasõnastiku abil lapse- ja hoidjakeele spetsiifilisemaks muuta. Hetkel on teada, et kindlasti tuleks kasutajasõnastikku täiendada pärisnimede ja reduplikatiivsete sõnade näol, kuid täpsema tegevusplaani jaoks peab tundmatuks jäänud ja vale analüüsi saanud sõnu põhjalikumalt analüüsima.

Spontaanse kõne lindistamine ja litereerimine on oluline ning töö- ja ajamahukas protsess, mistõttu tekib seda suurem vajadus ühtse transkribeerimistava järele. Artiklis oleme esitanud mõningaid soovitusi, kuidas korpuse märgendamist ja standardiseerimist morfoloogilise analüüsi jaoks kasulikumaks muuta (nt onomatopoeetilistele sõnadele, häälitsustele ja silpidele lisada spetsiaalseid märgendeid). Kõige olulisem ettepanek puudutab vigaseid vorme. CHILDES-keskkonnas veainfo esitamiseks tuleks kasutada märgendit [: *sihitud_sõna*], sest analüüsiprogramm on ülesehitatud nii, et ainult seda tüüpi vea esitamisel analüüsitakse nurksulgude vahel olevat sisu. Vea tähistamisel peab silmas pidama, et märgend järgneks vahetult igale vigasele sõnavormile, mitte tervele fraasile. Transkribeerimisel on alati olulised detailsus ja järjekindlus ning nii tekib ka automaatseks morfoloogiliseks märgendamiseks kvaliteetne sisend, mis võimaldab kvaliteetset väljundit. Olemasolev korpus on eesti lastekeele uurimiseks suurepärane andmestik ning selle lihtsamini märgendatavaks ja analüüsitavaks muutmine võimaldaks lastekeele kvantitatiivseid korpusuuringuid paremini teostada.

Viidatud kirjandus

Argus, Reili 2004. Imitatiivide kohast lastekeeles: reduplikatsioonist, morfoloogiast ja sõnaliigilisest ambivalentsusest. [Imitatives in child language: reduplication, morphology and vague word class distinctions.]– Eesti Rakenduslingvistika Ühingu aastaraamat, 1, 19–34. http://dx.doi.org/10.5128/ERYa1.01

Argus, Reili 2007. Eesti lastekeelekorpuse morfoloogilisest märgendamisest. [Morphological coding of Estonian child language database.] – Tallinna ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine. Tallinna Ülikooli eesti filoloogia osakonna toimetised 9. Tallinn: Tallinn Ülikooli Kirjastus, 65–86.

Argus, Reili 2008. Eesti lastekeelekorpuse morfoloogiliste vigade märgendamisest ja liigitamisest. [Coding and classification of morphological errors of Estonian child language database.] – Pille Eslon (Toim.), Õppijakeele analüüs: võimalused, probleemid, vajadused. Tallinna Ülikooli eesti filoloogia osakonna toimetised 10. Tallinn: Tallinn Ülikooli Kirjastus, 11–31.

Argus, Reili 2010. Mida teeb *tegema*-verb hoidjakeeles. [Constructions with the verb tegema 'do, make' in child directed speech.] – ESUKA / JEFUL, 1 (2), 17–34.

Argus, Reili; Kõrgesaar, Helen 2014. Sõnaliigid eesti lapse kõnes ja lapsele suunatud kõnes. [Word classes in the child's speech and in the child-directed speech.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 10, 37–53. http://dx.doi.org/10.5128/ERYa10.03

Aviad, Alber; MacWhinney, Brian; Nir, Bracha; Wintner, Shuly 2013. The Hebrew CHILDES corpus: transcription and morphological analysis. – Language Resources and Evaluation, 47 (4), 973–1005. https://doi.org/10.1007/s10579-012-9214-z

Behrens, Heike 2008. Corpora in language acquisition research: history, methods, perspectives. – Heike Behrens (Ed.), Corpora in Language Acquisition Research: History, Methods, Perspectives. Trends in Language Acquisition Research, 6. John Benjamins Publishing Company, xi–xxx. http://doi.org/10.1075/tilar.6

Behrens, Heike 2012. Corpus analysis of child language. – Heike Behrens, The Encyclopedia of Applied Linguistics. Blackwell Publishing Ltd, 1214–1222. https://doi.org/10.1002/9781405198431.wbeal0242

Burnard, Lou 2014. What is the Text Encoding Initiative? How to add intelligent markup to digital resources. http://books.openedition.org/oep/426 (11.5.2016). [Introduction; The TEI and XML; The structural organization of a TEI Document.]

CHAT. Codes of the Human Analysis of Transcripts. http://childes.psy.cmu.edu/manuals/CHAT.pdf (29.9.2016).

CHILDES. Child Language Exchange System. http://childes.psy.cmu.edu/data/ (29.9.2016).

CLAN. Computerized Language Analysis. http://childes.psy.cmu.edu/manuals/clan.pdf (29.9.2016).

EKK = Erelt, Mati; Erelt, Tiiu; Ross, Kristiina 2007. Eesti keele käsiraamat. [Handbook of Estonian.] Tallinn: Eesti Keele Sihtasutus.

Hennoste, Tiit 2000. Suulise eesti keele uurimine: transkriptsioon, taust ja korpus. [Research into spoken Estonian: transcription, background, corpus.] – Keel ja Kirjandus, 2, 91–106.

Hennoste, Tiit 2002. Suulise kõne uurimine ja sõnaliigi probleemid. – Renate Pajusalu, Ilona Tragel, Tiit Hennoste, Haldur Õim (Toim.), Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Tartu: Tartu Ülikool, 56–73.

Kaalep, Heiki-Jaan; Muischnek, Kadri; Müürisep, Kaili; Rääbis, Andriela; Habicht, Külli 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti keele testkorpuse morfosüntaktilise märgendamise kogemusest. [Do the available morphological descriptions of Estonian work on a real text?] – Keel ja Kirjandus, 9, 623–633.

Kaalep, Heiki-Jaan; Vaino, Tarmo 2000. Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis. – Tiit Hennoste (Toim.), Arvutuslingvistikalt inimesele. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1. Tartu: Tartu Ülikool, 87–101.

Kask, Paula Helena 2016. Eesti lastekeele andmete esitus andmepangas CHILDES. [Notation of Estonian child language in CHILDES-system.] Bakalaureusetöö. Tartu Ülikool, filosoofiateaduskond, eesti ja üldkeeleteaduse instituut. http://hdl.handle.net/10062/51867

Laing, Catherine E. 2014a. A phonological analysis of onomatopoeia in early word production. – First Language, 34 (5), 387–405. https://doi.org/10.1177/0142723714550110

Laing, Catherine E. 2014b. Phonological 'wildness' in early language development: Exploring the role of onomatopoeia. – Proceedings of the first Postgraduate and Academic Researchers in Linguistics at York (PARLAY 2013) conference.

MacWhinney, Brian; Snow, Catherine 1985. The child language data exchange system. – Journal of Child Language, 12 (2), 271–296. https://doi.org/10.1017/S0305000900006449

McEnery, Tony; Hardie, Andrew 2011. Corpus Linguistics: Method, Theory and Practice. Cambridge Textbooks in Linguistics. Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Muischnek, Kadri; Kaalep, Heiki-Jaan; Sirel, Raul 2011. Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile. [A corpus-based approach to the automatic morphological analysis of Estonian computer-mediated communication.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 7, 111–127. http://dx.doi.org/10.5128/ERYa7.07

Muischnek, Kadri; Orav, Heili; Kaalep, Heiki-Jaan; Õim, Haldur 2003. Eesti keele tehnoloogilised ressursid ja vahendid. Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Tallinn: Eesti Keele Sihtasutus

Muischnek, Kadri; Vider, Kadri 2005. Sõnaliigituse kitsaskohad eesti keele arvutianalüüsis. [The problems of word class disambiguation in the automatic analysis of Estonian.] – Eesti Rakenduslingvistika Ühingu aastaraamat, 1, 99–114. http://dx.doi.org/10.5128/ERYa1.05

Tomasello, Michael; Stahl, Daniel 2004. Sampling children's spontaneous speech: How much is enough? – Journal of Child Language, 31 (1), 101–121. https://doi.org/10.1017/S0305000903005944

Vaik, Kristiina 2016. Eesti lapsekeele korpuse morfoloogilisest märgendamisest. [Morphological annotation of the Estonian child language corpus.] Magistritöö. Tartu Ülikool, filosoofiateaduskond, eesti ja üldkeeleteaduse instituut. http://hdl.handle.net/10062/52841

Vihman, Virve-Anneli 2015. Pick it up: A look at referential devices in Estonian Child-Directed Speech. – ESUKA / JEFUL, 6 (2), 63–83. http://dx.doi.org/10.12697/jeful.2015.6.2.03

Kristiina Vaigu (Tartu Ülikooli arvutiteaduse instituut) uurimisvaldkondadeks on morfoloogia, korpuslingvistika ja arvutilingvistika.
J. Liivi 2-303, 50409 Tartu, Estonia
kristiina.vaik@ut.ee

Virve-Anneli Vihmani (Tartu Ülikool) teadustöö põhisuundadeks on eesti keele morfosüntaks, laste morfosüntaktiline areng ning keelepoliitika.
Jakobi 2, 51014 Tartu, Estonia
virve.vihman@ut.ee

ISSUES IN MORPHOLOGICAL ANNOTATION OF THE ESTONIAN CHILD LANGUAGE CORPUS

Kristiina Vaik, Virve-Anneli Vihman
University of Tartu

This article presents the results of an initial attempt to automatically annotate the currently existing, publicly available Estonian child language corpus morphologically. CLAN software is not suitable for morphological analysis of Estonian, but Estonian language technology resources are available for written language and can be adapted to spoken language and specific genres. The automatic parser provided annotation for 92–98% of words in the child-directed speech and 57–96% of the child speech, with the results for child speech varying across corpora. A manual analysis was also conducted of words which were automatically annotated in a random selection of transcriptions from each corpus. Across corpora, 63–96% of annotated words were correctly annotated. Reasons for the variation are discussed, and obstacles to automatic annotation are identified at various levels.

First, the corpora have been collected and transcribed with various goals and according to differing principles, hence the style and detail of transcription vary greatly across the corpora. Second, even within a single corpus, discrepancies appear in coding which need to be uniformly resolved in order to ensure accurate morphological annotation. Finally, for flagging non-standard or idiosyncratic forms, the implementation of metacodes available for use in the child language corpora would greatly assist the task of automatic morphological parsing. For each corpus, a user dictionary adapted to the particular genre and the particular corpus would need to be developed, including proper names and idiosyncratic words. The marking of errors is a crucial area which needs to be standardised in order to enable automatic annotation. Additionally, five groups of words which received inaccurate annotation were identified, and suggestions are made for transcription of child language corpora in order to ease the task of morphological annotation in the future.

Keywords: child language, corpus, automatic annotation, transcription, Estonian