TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Paula Helena Kask

# Eesti lastekeele andmete esitus andmepangas CHILDES

Bakalaureusetöö

Juhendaja PhD Sirli Zupping

Tartu 2016

## Sisukord

## Sissejuhatus

Bakalaureusetöös käsitletakse rahvusvahelise lastekeele andmepanga CHILDES eesti lastekeelekorpuse ülesehitust ja transkribeerimisprobleeme. Töö teema paigutub keeleomandamise ja korpuslingvistika valdkonda.

Laste kõne arengu uurimused põhinevad enamasti kas eksperimentidest saadud andmetel või korpustes sisalduval keelematerjalil. Lisaks lingvistidele uurivad lastekeelt mitmed teisedki teadlased, kes oma erialal puutuvad rohkemal või vähemal määral kokku kõne arengu jälgimisega, näiteks psühholoogid, kasvatusteadlased, neurofüsioloogid ja defektoloogid. (Argus 2003: 27)

Siinses töös keskendutakse laste spontaanse kõne materjali kogumise ja talletamise ühele võimalusele, milleks on lastekeelekorpus. Kõige rohkem eesti lastekeele suulisi tekste on hetkel võimalik leida andmekogust CHILDES (Child Language Data Exchange System)[1]. See on rahvusvaheline, mahukas ja usaldusväärne andmekogu, mille suurimaks väärtuseks võib pidada andmete esitamise ühtlustatust (Argus 2007: 66). CHILDES on elektrooniline keeleressurss, milles paiknevad lastekeelekorpused on üldiselt kasutatavad vabavarana.

CHILDESi eesti alamkorpuste koostajateks on siiani olnud lastekeeleuurijad, kes on spontaanse kõne andmeid lindistanud ja litereerinud oma uurimuste tarbeks. Samuti on eesti alamkorpused kodeeritud uurija huve silmas pidades (Argus 2007: 66). Litereerimisel on üldiselt lähtutud kuuldeortograafia põhimõttest, st sõnad pannakse kirja nii, nagu neid hääldatakse (Argus 2007: 71). CHILDES-süsteemis kasutatav litereerimisformaat on andmete sisestaja jaoks üsna paindlik ja mugav ning ei tee rangeid ettekirjutusi transkribeerimisreeglite kohta. See loob olukorra, kus litereerija saab tekste kirja panna n-ö oma kohandatud süsteemi järgi, mis võib põhjustada litereeritud tekstide ebaühtluse. Reili Argus (2007) on viidanud samale probleemile ning osutanud, et tulevastele keeleuurijatele mõeldes tasuks see andmekogu tähelepanelikult

[1] Korpus on leitav internetileheküljel http://childes.psy.cmu.edu (19.05.2016).

üle vaadata ning ühtlustada, sest kõik transkribeerimistasandil tehtavad otsused mõjutavad analüüsitulemusi (2007: 77, 84).

Eelnevast lähtuvalt on käesoleva töö üldine eesmärk kaardistada CHILDESi eesti lastekeelekorpuse hetkeseis. Alleesmärk on uurimise käigus välja selgitada, millised erinevused on alamkorpustes rakendatud litereerimispõhimõtetel ning millest need tulenevad. Lisaks antakse uurimistulemuste põhjal soovitusi edasiseks tööks CHILDESi eesti alamkorpuste koostamisel.

CHILDESi eesti lastekeelekorpuse kohta pole varem sellist uurimust tehtud, kus kaardistatakse korpuse hetkeseis ja võrreldakse litereeringute teksti. Siiski on lastekeelekorpuse koostamispõhimõtetele osutatud Reili Arguse artiklites (2007, 2008).

Siinses uurimuses on kasutatud kvalitatiivset lähenemist: kaardistatakse hetkeseis, analüüsitakse korpuse materjali ning uuritakse, millised on korpusekoostajate probleemid seoses litereerimisega. Töö empiiriline osa põhineb kaht liiki andmetel: CHILDESi eesti alamkorpuste litereeringutel[2] ja korpusekoostajate küsitlusel. Eesti lastekeelekorpuse alamkorpustes sisalduva materjali ülevaates tutvusin kõikide andmetega, mis on esitatud lindistuste päises: lindistuse aeg, kestus, situatsioon, osalejad jne. Täpsemate andmete kogumiseks (st transkriptsioonimärgid, sõnade ülesmärkimise viisid jm) valisin igast korpusest kolm lindistust: lindistusperioodi algusest, keskelt ja lõpust. Juhul, kui ühes korpuses oli lindistatud rohkem kui ühe lapse kõnet, valisin iga lapse kohta ühe lindistuse. Selleks, et saada täpsemat informatsiooni iga alamkorpuse kohta, koostasin lühikese küsimustiku, mille edastasin kõigile korpusekoostajaile.

Bakalaureusetöö koosneb viiest osast. Töö teoreetilise osa esimeses peatükis antakse ülevaade suulisest kõnest, selle kogumise ja talletamise viisist ning litereerimisest. Teine peatükk keskendub lastekeele mõistele ja uurimismeetoditele. Kolmas peatükk avab töö empiirilise osa: kirjeldatakse andmepanka CHILDES ja selles sisalduvat eesti lastekeelekorpust. Sama peatüki lõpetab ülevaade küsitluse tulemustest, mis on omakorda sissejuhatuseks eesti lastekeelekorpuse andmete analüüsile. Seega, neljandas

[2] Edaspidi on samas kontekstis kasutatud ka terminit „lindistus“, sest alamkorpuste struktuur on üles ehitatud lindistuste kaupa.

peatükis on esitatud kokkuvõtlikult ja süsteemselt eesti lastekeelekorpuse alamkorpuste litereeritud tekstide erinevused ja sarnasused. Töö viies peatükk võtab kokku analüüsi tulemusel selgunud transkribeerimise erinevused ning seal antakse soovitusi edasiseks tööks CHILDESi eesti lastekeelekorpuse koostamisel.

# 1. Suuline kõne

Esimese keele omandamise protsess kulgeb enamasti suulise kõne vahendusel. Selle kõrval on olulised ka muud väljendusvormid, näiteks žestid ja viiped, kuid keskseks saab pidada siiski keelt ja kõnet, mida laps kuuleb ning kasutab.

Inimkeel erineb teiste liikide suhtlusest kahe põhilise tunnuse poolest: esimene neist on see, et inimeste keeleline suhtlus on sümboolne. Inimesed kasutavad suheldes keelelisi sümboleid, et anda edasi infot. Teine põhiline eristav tunnus on see, et inimeste keeleline suhtlus on grammatiline. Inimesed kasutavad suheldes keelelisi sümboleid mustriliselt – neid mustreid teame kui keelelisi konstruktsioone, millel on oma tähendus. (Tomasello 2003: 8)

Keeleuurimine koosneb mitmest keeleteaduse osast, keeleteaduses on kogum uurimisvaldkondi, teemasid ja lähenemisi. Esimene suurem rühm koondab endasse tähenduse uurimise, teine alarühm diskursuse- ja tekstianalüüsi, kolmanda alarühma moodustab sotsiolingvistika. (Hennoste 2002: 217)

Lisaks eelnevale on veel oluline liigendus kahe suure metaregistri alusel: kirjaliku keele analüüs ning suulise keelekasutuse ehk kõnekeele uurimine (Hennoste 2002: 218). Siinses uurimuses keskendutakse suulisele kõnele.

Suuline kõne ja kirjaliku keele üks olulisemaid erinevusi on see, et suulises kõnes ei kehti kirjakeele normingud (Hennoste 2000a: 2235). Kirjalikus keeles liigituvad üksused lauseteks, suulises keeles aga ei ole lause keskne üksus, sest suulises kõnes on parandused, takerdumised ning tugev kontekstuaalsus (Hennoste 2000a: 2223). Kui püüda suulise kõne elemente kuidagi süstematiseerida, siis võib välja tuua suulise kõne kesksete üksuste kolmese jaotuse: semantilis-intonatsioonilised üksused, lausungid, kõnevoor ja paratoon (Hennoste 2000a: 2223).

Esimese liigenduse, semantilis-intonatsiooniliste üksuste alla kuuluvad toonigrupp, süntagma ja ideeüksus (Hennoste 2000a: 2223), mida vaadeldakse lähtuvalt sellest, et suuline spontaanne kõne on purskeline ja portsjoniline (Hennoste 2000a: 2224). Kõige olulisemaga, ideeüksusega väljendab kõneleja fookustatud mõtet keeleliselt. Ideeüksuse

piiriks on kas hääletoon või takerdumine. Hääletooni muutumine näitab idee lõppemist, kuid seejuures tekst jätkub. Mõtte takerdumisele viitavad pausid, kordused, üneemid (näiteks *ee, ää, õõ*) ja partiklid. (Hennoste 2000a: 2224)

Vene lingvisti Lev Štšerba järgi (Hennoste 2000a: 2224 kaudu) koosneb süntagma sõnast, sõnaühendist või nende grupist, süntagma väljendab kõneprotsessis mõttelist tervikut.

Toonigrupiks nimetatakse kõneleja viisi teksti organiseerimisel. Kõik toonigrupid annavad edasi mingisugust ühtset osa informatsioonist. Seega toonigrupi eesmärk on anda kuulajale edasi terviklikku teavet. (Hennoste 2000a: 2225)

Kui kirjaliku keele üksused on laused, siis suulises kõnes kõige sarnasemad üksused lausetele on lausungid (Hennoste 2000a: 2226). Lindistatud kõne kuulamisel on võimalik eristada üksusi, mis lõppevad tooni langemisega, kusjuures enamasti viitab tooni langemine lausungi lõpetamisele. Keeleuurijad liigendavad niisuguste üksuste abil suulist vestlust. Lisaks intonatsioonile aitavad kõnet üksusteks jaotada ka pausid ja takerdumised. Oluline on märkida, et suulise kõne üksus ei pea olema terviklik lause ega hoopiski mitte lausesarnane üksus. Suulises kõnes võib üksuseks olla elliptiline lause, fraas, sõna või üneem. (Hennoste 2000a: 2227)

Vestluse või dialoogi põhiüksus on kõnevoor. Kui vahetub kõneleja, toimub ka vooruvahetus. Oma sisult on voor see, kui üks kõneleja on jätkuvalt hääles. Enamasti on voorude vahetumine sujuv. (Hennoste 2000a: 2229)

Vooruvahetuseks on neli võimalust. See võib olla sujuv, toimuda otsarääkimisel, vahetuda pausiga või pealerääkimisega (Hennoste 2000a: 2229). Otsarääkimine toimub, kui kõneleja ütleb oma vooru alguse lõppeva vooru otsa, seejuures ilma pausita (Hennoste 2000a: 2229), pealerääkimine tähendab aga seda, et kuulaja alustab enne kõneleja lõpetamist (Hennoste 2000a: 2230).

Suulise kõne litereerimisel või selle tekstide analüüsimisel on oluline teada, et kõnevooruks võib olla lause või mitu lauset, fraas, sõna, häälitsus, naer jne (Hennoste 2000a: 2230). Hennoste (2000: 2232-2233) järgi tähendab see, et ka „ühemorfeemilisi lühikesi häälesolekuid“ peetakse kõnevooruks, näiteks üneemid ja interjektsioonid (*ahah, mhmh* jne).

Suulist kõnet ei ole võimalik uurida enesevaatluse või juhuslike tähelepanekute põhjal (Hennoste 2002: 238). Selleks, et suulist kõnet saaks uurida, tuleb vastav materjal kuhugi talletada, tänapäevaseim lahendus on järgmine: "suulise kõne analüüs nõuab korpust" (Hennoste jt 2009: 111).

„Korpus on loomuliku keele tekstide kogu, mis on koostatud iseloomustamaks keele hetkeseisu või muutumist" (Muischnek jt 2003: 9). Keelekorpuseks nimetatakse korpust, mis sisaldab andmeid kirjalikust või suulisest keelest. Arvutiajastul tähendab korpus peamiselt elektroonilist tekstikogu. (Muischnek jt 2003: 9)

Kuna suulise kõne uurimine on korpusekeskne kogu maailmas, on korpuse koostamine oluline. Seejuures on omakorda tähtis transkriptsiooni valik ja suhtlussituatsiooni kirjeldav taustamaterjal (Hennoste 2002: 238).

## 1.2. Suulise kõne litereerimine eesti keeles

Igasugune keelekasutus on pidevas muutumises ja seda mõjutavad erinevad tegurid. Keelekasutust mõjutavad konkreetset inimest puudutavad omadused nagu haridus, aga ka olukorrast sõltuvad, näiteks mõjutab keelt see, kas tegemist on avaliku või privaatse vestlusega. Kuna keelelisi varieeruvusi on nii haruldasi kui ka neid, mida esineb rohkem, on suhtluse uurimine õnnestunum, kui on olemas korpus, mis erinevaid allkeeli sisaldab. (Hennoste jt 2009: 112) Näiteks, Tartu ülikoolis on loodud suulise eesti keele korpus, mis on mõeldud eeskätt keeleuurijaile (Hennoste jt 2009: 111).[3]

Tartu ülikooli suulise kõne korpuses on salvestatud ja transkribeeritud materjal. Korpuses on audiosalvestused ja videosalvestused, materjali salvestatakse digitaalselt. Selleks, et materjali analüüsida, on tarvis see transkribeerida. Transkribeerimiseks kasutatakse programme Praat või CLAN. (Hennoste jt 2009: 113) Eesti lastekeelekorpusse (lähemalt selle kohta 3. ptk-s) sisestatud materjal koosneb samuti litereeringufailidest, mis on loodud programmiga CLAN. Lastekeele kõnesalvestusi on püütud ka automaatselt transkribeerida, kuid seni on see jäänud vaid katsetamiseks (Maarits 2011).

[3] Tartu Ülikooli suulise kõne uurimisrühma koduleht http://www.cl.ut.ee/suuline/ (19.05.2016).

Tartu Ülikooli suulise kõne uurimisrühm on oma töös lähtunud muu hulgas sellest, et kõnesalvestuste transkribeerimisel märgitakse ära suhtlusüksused, sõnad ja suhtlushäälitsused (*ee, mhmh*), pausid, kõne prosoodilised ja paralingvistilised omadused (venitused, katkestused jmt), vestlusosaliste teineteisele peale- ja otsarääkimised, transkribeerija kahtlused (halvasti kuuldud sõnad), nähtuste kirjeldused (kõrvalised hääled jmt). (Hennoste jt 2009: 113−114)[4]

[4] Ülevaade TÜ suulise kõne korpuses kasutatud transkriptsioonimärkidest on leitav aadressil http://www.cl.ut.ee/suuline/Transk.php (19.05.2016).

## 2. Lastekeele uurimisest

Järgnevas peatükis keskendutakse lastekeele spontaanse kõne uurimismeetoditele. Sellele eelneb ülevaade lastekeele mõistest ja olemusest.

### 2.1. Lastekeele taust

Lastekeel on keel, mida räägivad lapsed, see on suuline spontaanne kõne. Lastekeele juurde kuulub ka mõiste hoidjakeel ehk lastele suunatud keel, mille erilisteks tunnusteks on näiteks kõrgenenud hääletoon või sagedased kordused (Clark 2009: 32, 36).

Lapse keelelise arengu motiveeritus lähtub kahest aspektist: soov suhelda teistega ning tahe olla ümbritsevate inimeste moodi, ehk neid järele teha. Kõnelema hakates on lastel vaja leida tähendused sõnadele ja väljenditele ning leida parim suhtlemise viis, et oma kavatsusi teistele selgitada. (Tomasello 2003: 31)

Nende probleemide lahendamiseks peavad nad avastama, kuidas kohandada oma lauseid iga vastuvõtja jaoks. Probleemide lahendamiseks on vaja järjekindlust, et õppida erinevates olukordades kasutama kindlaid lausungeid. (Clark 2009: 75)

### 2.2. Lastekeele uurimismeetodid

Lastekeele uurimist võib liigitada keeleandmete kogumise meetodite järgi ning lapse arenguperioodi alusel. Keeleandmete kogumise meetodite põhjal on uurimisperioodid jaotatud kolmeks, st uurimismeetodite arengu ja kasutuselevõtu aja järgi; lapse kõne arenguperioodi saab väga üldistavalt jagada kaheks, st aeg enne kõnelema hakkamist ja kõne arengu periood üldiselt.

Esimese keele omandamise uurimismeetodid jagunevad suures osas kaheks: kõne-eelsel perioodil kasutatavad meetodid ja kõne arengu perioodil kasutatavad uurimismeetodid (Parm 2013: 11). Siinse töö keskmes on korpusesse kogutud spontaanse kõne materjal, mis paigutub eeltoodud jaotuse järgi kõne uurimise meetodite alla.

Lastekeele uurimise meetodite kasutuselevõtt on seotud üldise tehnoloogilise arenguga, samuti keeleomandamise distsipliini arenguga. Keeleandmete kogumise meetodi põhjal on lastekeele uurimisperioodid jagatud kolmeks. Esimene periood põhineb päevikumärkmetel ning see oli valdav aastatel 1876–1926. (Argus 2003: 27) Sel ajal kuulus lapse keelelise arengu jälgimine väikese osana lapse üldise arengu jälgimise juurde. Uurijateks olid lapsevanemad ise, muu hulgas tegid nad tähelepanekuid lapse keele omandamise kohta. (Argus 2003: 28)

Teine periood oli aastatel 1926–1957 (Argus 2003: 27), mil eesmärgiks oli jälgida samal ajal suure hulga laste keele arengut. Last vaadeldi kui passiivset olendit, kes oli keskkonna poolt kontrollitud. Valim koosnes tavaliselt 100−200 lapsest ning üritati määrata lapse normaalne keeleline käitumine. (Argus 2003: 28)

Kolmanda perioodi meetod on lastekeele uurimises valdav ka praegu. See periood sai alguse 1957. aastast (Argus 2003: 27) ning seda nimetatakse pikaajaliste uuringute perioodiks. Selle perioodi nimi tuleneb uurimismeetodist ning võib öelda, et see koosneb kahest eelnevast perioodist. Nimelt jälgitakse lapsi pikema ajaperioodi vältel ning uuritavad lapsed valitakse kindlate kriteeriumite põhjal (Argus 2003: 28). Lindistatakse 30–45 minuti pikkuseid dialooge ning vaadeldavate laste arv on väiksem. Sellele meetodile on omane, et lapsi saab jälgida põhjalikult. (Argus 2003: 29)

Lisaks pikiuuringutele on viimasel ajal kerkinud esile ka eksperimentaalsete meetodite kasutus. Katsetüüpe on erinevaid ning need sõltuvad sellest, millised on uurija eesmärgid. Näiteks võib katse sisuks olla reaktsioonikiiruse mõõtmine, imiteerimisülesanne, mõistmistestid, loomistestid jne. (Parm 2013: 12)

Spontaanse kõne andmeid hakatakse üldjuhul koguma siis, kui lapse kõnesse tekib esimene sõna, pikiuuringu jaoks kogutakse andmeid enne lapse nelja-aastaseks saamist ehk kõne arengu varasel perioodil. Spontaanse kõne andmeid võib koguda kirjalike märkmete abil või audio- ja videolindistusi tehes. (Parm 2013: 11)

Selleks, et kogutud keeleandmeid paremini analüüsida, on oluline, et need oleksid kogutud elektroonilisse korpusesse. Sellisel juhul on saadud andmete automaattöötlus, kõrvutamine ja võrdlemine keeleuurijale lihtsam. (Parm 2013: 11)

# 3. Andmepank CHILDES ja CHILDESi eesti lastekeelekorpus

Järgnevas peatükis tutvustatakse CHILDESit ja selle eesti lastekeelekorpust. Antakse ülevaade CHILDESist ning sellele järgneb eesti lastekeelekorpuse alapeatükk, mis tutvustab eesti alamkorpusi. Lisaks keskendutakse siinses peatükis korpuse koostamise ja andmete litereerimise põhimõtetele.

## 3.1. CHILDESi tutvustus

Andmepank CHILDES loodi 1984. aastal USAs Brian MacWhinney ja Catherine Snow eestvedamisel (MacWhinney 2001a: 2). Andmepangas on korpusi nii esimese kui ka teise keele omandamisest. Lisaks leidub selles neid korpusi, mille eesmärk on abistada logopeediliste probleemide ning kakskeelsuse uurijaid.

CHILDESis on transkribeeringuid rohkem kui 40 eri projektist inglise keelest ning lisaks veel 20 keele materjali: eesti, mandariini-hiina, kantoneesi-hiina, taani, hollandi, prantsuse, saksa, kreeka, heebrea, ungari, itaalia, jaapani, mambila, poola, portugali, vene, rootsi, tamili, türgi ja ukraina (MacWhinney 1995: 1).

CHILDESit võib pidada usaldusväärseks keeleandmete esitamise ühtluse poolest, mis tagab, et erinevate keelte uurijad saavad uurimistulemusi võrrelda (Argus 2007: 66).

Kuna keeleandmete kogumine, transkribeerimine ja analüüsimine on ajakulukas, arendas CHILDESi meeskond välja vahendid, mis olukorda parandaksid. Vahendid loodi selleks, et andmete jagamine oleks hõlpsam, transkriptsioon usaldusväärsem ning andmete analüüs automaatne. (MacWhinney 2001b: 1)

Laias laastus koosnebki CHILDES-süsteem kolmest osast: CHAT, CLAN ja transkribeeritud keelte andmebaas, mis on koostatud programmiga CLAN CHAT-formaadis (MacWhinney 2001b: 3).

CLAN (Child Language Analysis) on programm, mis on loodud keele analüüsimiseks. See toetab viit põhilist keeleanalüüsi tüüpi: leksikaalne analüüs, morfoloogiline analüüs, süntaktiline analüüs, diskursusanalüüs ning fonoloogiline analüüs (MacWhinney 2001b: 12).

Transkriptsioonisüsteemi CHAT kasutatakse andmete töötlemiseks. Transkribeeritud keelematerjalile on lisatud päis, mis annab uurijale teavet lindistuse aja, koha, osalejate, kestuse, lapse vanuse, lindistussituatsiooni ning vastavalt vajadusele ka teiste asjaolude kohta. CHAT-formaadis tähistatakse kõneleja kolmetähelise koodiga, mis on paigutatud põhiridadele (nt lapse kohta üldine kood *CHI või nime eesmistest tähtedest koosnev kombinatsioon). Sellele järgneb vestluses osalejate tegelik kõne, sõltridadele lindistaja, ning transkribeerija ja uurija kommentaarid või kodeeringud. (Argus 2007: 68)

## 3.2. CHILDESi eesti lastekeelekorpus

Alates 1998. aastast sisaldab CHILDESi korpus ka eesti lastekeele andmeid. Eesti lastekeelekorpuse alamkorpusi nimetatakse nende koostajate järgi. 2016. aasta seisuga on CHILDESis eesti lastekeele alamkorpusi seitse: Argus, Beek, Kapanen, Kohler, Kõrgesaar, Vija ja Zupping.

Materjali on CHILDESi eesti lastekeelekorpuses kokku umbkaudu 169 tundi. Lapsi on lindistatud väga erinevas vanuses: kõige noorem on Beeki alamkorpuses Liisbet, kes oli esimese lindistuse ajal vanuses 0;9.24[5]. Kõige vanem lindistatud laps on Kõrgesaare alamkorpuses Harley vanuses 14;1.0.

Kokku on alamkorpustes 25 lapse materjal, millest poiste alamkorpusi on 11: Hendrik, Andreas, Carlos, Henri, Sandor, Taimo, Gregory, Harley, Ruuben, Andry, Artur. Tüdrukute alamkorpusi on 14: Liisbet, Linda, Martina, Anna, Helen, Mari, Stella, Kaisa, Arabella, Hellyn, Jaana, Mia, Olivia, Sirlin. Ülevaade eesti lastekeelekorpuse andmetest on esitatud tabelis 1.

Enamasti on materjali lindistatud kodustes tingimustes, näiteks lapse kodus või ka lindistaja juures külas. Vestlusi peetakse igapäevatoiminguid tehes ja mängides.

Eesti lastekeelekorpuse dialoogides osalevad laps, lapsevanemad, vanavanemad, õed-vennad, mängukaaslased ja lindistaja. Kõige rohkem on dialooge ema ja lapse vahel, kuid esineb ka lindistusi, kus osalevad ema, laps ning lindistaja.

[5] CHILDESis märgitakse lapse vanust järgmiselt: aasta;kuu.päev.

Järgnevalt on esitatud eesti lastekeelekorpuse alamkorpused nende koostajate nimede järgi. Korpuse nimele järgneb lühidalt teave alamkorpuses sisalduva kohta.

Arguse alamkorpuses on lindistatud Hendriku kõnet, mil laps on vanuses 1;8.13 kuni 2;5.30. Korpuses on andmed igapäevaelu vestlustest, milles osalevad Hendrik, tema vanemad, vend ning mõnikord ka vanaema. Situatsioonid on kodused. Lindistusi on selles korpuses 17.

Beeki alamkorpuses on Liisbeti kõne. Liisbeti lindistused on tehtud vanuses 0;9.24 kuni 2;5.0. Lindistustes osalevad laps, ema ja isa. Lindistused on läbi viidud kodustes situatsioonides: mängimine, potitamine ja söömine. See alamkorpus sisaldab 20 lindistust.

Kapaneni alamkorpuses on lindistatud Martinat vanuses 1;3.15 kuni 3;1.0. Neis lindistustes osalevad Martina, tema isa ja ema. Vestlused on sageli tehtud hommikusöögilauas või teiste hommikuste toimingute ajal. Selles alamkorpuses on 11 lindistust.

Kohleri alamkorpuses on 8 lapse lindistused: Anna, Carlos, Helen, Henri, Mari, Sandor, Stella ja Taimo. Annat on lindistatud vanuses 1;11.20 kuni 2;0.17. Lindistusi on Anna korpuses 7. Neis osalevad Anna, tema mängukaaslane, lapsehoidja ning lindistaja**.** Need lindistused on tehtud lapsehoidja juures ning liivakastis.

Carlosega on tehtud 9 lindistust vanuses 1;7.17 kuni 1;10.29. Need lindistused on tehtud lastetoas, lapsehoidja juures ning kodus. Lindistustes osaleb Carlos, tema ema, Carlose mängukaaslane, õde ning lindistaja.

Helenit on lindistatud vanuses 1;1.17 kuni 1;10.17. Heleni korpuses on lindistusi 7, neis osalevad ema, isa, laps ja lindistaja. Lindistussituatsioonid on kodused.

Henri korpuses on 3 lindistust. Neis osalevad Henri, tema vend, vanaema ning lindistaja. Kõik need lindistused on tehtud vanaema juures, räägitakse mänguautodest ja teistest mänguasjadest. Henrit on lindistatud vanuses 2;2.12 kuni 2;3.08.

Mari korpuses on 7 lindistust. Need on tehtud ajavahemikus, mil Mari oli vanuses 2;5.07 kuni 2;8.10. Neis osalevad laps, ema, vend, mängukaaslane ja uurija. Lindistussituatsioonid on kodused ja neis toimuvad igapäevategevused ja -mängud.

Sandori korpuses on 10 lindistust, mis on tehtud vahemikus, kui Sandor oli vanuses 1;2 kuni 2;2.22. Lindistustes osalevad Sandor, tema vanemad ning lindistaja. Lindistused on läbi viidud lapse kodus ja lindistaja kodus.

Stella korpuses on lindistusi 9. Lindistustel on Stella vanuses 0;11.22 kuni 1;6.04, neis osalevad laps, tema vanemad, vend ning lindistaja. Lindistused on läbi viidud kodus, elutoas.

Taimo korpuses on samuti 9 lindistust. Osalejateks on Taimo, tema ema ja isa ning lindistaja. Lindistused on tehtud, kui Taimo oli vanuses 1;5.08 kuni 1;11.13. Lindistuste toimumisohaks on lapse kodu.

Kõrgesaare alamkorpuses leiab 12 korpust, mille nimed on Gregory, Harley, Kaisa, Ruuben, Andri, Arabella, Artur, Hellyn, Jaana, Mia, Olivia ja Sirlin. See korpus on loodud eelkõige hoidjakeele uurimise eesmärgil.

Lisaks on Kõrgesaare korpuses kaks täiskavanutevahelist dialoogi, milles vanemad arutavad remondiplaane.

Gregory korpuses on 10 lindistust tehtud ajavahemikus, kui Gregory oli vanuses 6;6.19 kuni 10;5.11. Lindistustes osalevad Gregory, tema ema-isa ja Gregory vend. Lindistustes toimuvad tegevused on näiteks meisterdamine, kodutööde tegemine ja lauamängu mängimine.

Harley korpuses on 21 lindistust. Lindistused on tehtud, kui Harley oli vanuses 4;0.21 kuni 14;1.0. Neis osalevad laps, lapsevanemad ja lindistaja. Lindistustes vestleb laps vanematega erinevatel teemadel, näiteks loomadest või koolist. Juuakse ka teed ja vaadatakse aabitsat.

Kaisa korpuses on lindistusi 2. Kaisa vanus on lindistusperioodil 5;8.12 ja 5;9.4. Lindistustes osalevad ema, laps ja isa. Lindistussituatsioonides on tegevusteks koristamine, söögivalmistamine, joonistamine ja mängimine.

Ruubeni korpuses on lindistusi 4. Neis osalevad Ruuben ning tema vanemad. Lindistustel on Ruuben vanuses 1;3.3 kuni 3;6.2. Lindistustel on isa ja laps külas.

Andriga on tehtud kaks lindistust, millest esimesel on ta vanuses 11;7.17 ning teisel 11;9.7. Lindistusel arutavad Andri ja lindistaja maailmaasju. Teisel lindistusel osalevad Andri, lindistaja ja õde. Kaetakse ema sünnipäevalauda.

Järgnevad lindistused ei ole eraldi korpustena esitatud. Iga lapse kohta on üks lindistus, mis on lisatud Kõrgesaare korpusesse. Arabella on lindistusel vanuses 1;8.6. Lindistuses osalevad Arabella ja tema isa, kes on külla läinud.

Artur on vanuses 1;4.15. Vestluses osalevad laps ja lindistaja, tehakse igapäevaseid toimetusi lapsega: söömine ja potitamine, mängimine.

Hellynit on lindistatud vanuses 8;7.13. Lindistuses osalevad ema ja laps, kes mängivad lauamängu. Jaana lindistusel osalevad ema ja Jaana. Laps on vanuses 2;5.12. Lindistusel mängitakse puslega.

Mia lindistusel on Mia ja ema. Mia on vanuses 2;3.17. Mia ja ema panevad koos puslet kokku ja mängivad klotsidega.

Olivia lindistuses on lisaks Oliviale ema, vend ja lindistaja. Olivia vanus on 3;2.12. Laps sööb kommi ja täiskasvanud ajavad juttu.

Sirlini lindistuses on Sirlin, tema ema ja lindistaja. Sirlini vanus on 1;3.17. Lindistaja ja laps mängivad põrandal ja ema räägib lapsega.

Vija korpuses on lindistused Andreasega vanuses 1;7.24 kuni 3;1.13. Lindistused on kodused situatsioonid: söömine, mängimine, õhtused ja hommikused toimetused, esineb ka olukordi, kus laps uurib lindistussüsteemi. Vestlustes osalevad laps, lapsevanemad ja vahel ka vanaema.

Zuppingu korpuses on lindistatud Lindat vanuses 1;3.3 kuni 4;2.13. Lindistustes osalevad Linda, tema ema ja vend. Situatsioonid on enamasti kodused toimetused ja mängud. Linda korpuses on ka lindistusi, mis on üles võetud väljaspool kodu. Lindistusi on Zuppingu alamkorpuses 23.

Tabel 1. **CHILDESi eesti lastekeelekorpuse üldandmed (seisuga mai 2016).**

| Alamkorpuse nimi | Korpuse nimi | Lindistatud laste arv | Lindistatud laste vanus | Lindistuste arv alamkorpuses |
|---|---|---|---|---|
| Argus | Hendrik | 1 | 1;8.13–2;5.30 | 17 |
| Beek | Liisbet | 1 | 0;9.24–2;5.0 | 20 |
| Kapanen | Martina | 1 | 1;3.15–3;1.0 | 11 |
| Kohler | Anna, Carlos, Helen, Henri, Mari, Sandor, Stella, Taimo | 8 | 0;11.22–2;8.10 | 61 |
| Kõrgesaar | Gregory, Harley, Kaisa, Ruuben, Andri, Arabella, Artur, Hellyn, Jaana, Mia, Olivia, Sirlin | 12 | 1;3.3–14;1.0 | 48 |
| Vija | Andreas | 1 | 1;7.24–3;1.13 | 74 |
| Zupping | Linda | 1 | 1;3.3–4;2.13 | 23 |

Eesti lastekeelekorpus täieneb jooksvalt seoses riikliku programmi „Eesti keel ja kultuurimälu“ projektiga „Eesti laste- ja hoidjakeele korpuse täiendamine ja kaasajastamine“, mille eesmärk on kasvatada korpuse lindistuste maht 300 tunnini.[6]

## 3.4. Korpuse koostamispõhimõtted

Keelematerjali kodeerija peab transkribeerides otsustama, kas mingi keeleline üksus on normile vastav või mitte, samuti peab otsustama, kuidas tähistada normist erinevat keeleüksust. Nende otsuste langetamise järjekindlus määrab kodeeringu usaldusväärsuse uurija jaoks. (Argus 2008: 21)

Praegu esineb veakodeeringus erinevusi: näiteks on põhireal kasutatud selgituse ehk tõlke lisamist, kasutades erinevaid tähistusi (võrdusmärki ja koolonit) ning “mõnede vigade tarbeks on kasutatud eraldi vigade sõltrida” (Argus 2008: 22). Sellest täpsemalt käesoleva töö 4. peatükis.

[6] Vt Eesti Teadusinfosüsteemi koduleht: https://www.etis.ee/Portal/Projects/Display/5d01b766-7ca4-4af3-b498-570b4d1e9909 (24.05.2016).

### 3.4.1. Koostamispõhimõtted üldiselt

Arvutipõhise keeleandmete süsteemi puhul on kolm olulist eesmärki: selgus (MacWhinney 2000: 14), loetavus ja andmete sisestamise lihtsus (MacWhinney 2000: 15). Selgus tähendab seda, et kodeerides peab igal sümbolil olema selge vaste, mida saab siduda mõne sõnaga. Oluline on silmas pidada ka süsteemsust. (MacWhinney 2000: 14) Igal koodil peaks alati olema unikaalne tähendus, mis ei sõltu teistest koodidest (MacWhinney 2000: 15).

Nii nagu inimeste keelt peab olema lihtne töödelda, peaks transkribeeritud teksti olema lihtne lugeda. CHILDES-süsteemi loojad on püüdnud tagada CHATi kasutamisel mitmeid valikuvõimalusi, mille abil kasutaja saab transkribeeringu võimalikult hõlpsasti loetavaks muuta. (MacWhinney 2000: 15)

Andmete sisestamise lihtsuse printsiibi puhul kehtib CLANi programmis põhimõte, et transkribeerijale pakutakse andmete sisestamisel arvutipõhist abi. Näiteks andmete täpsuse automaatkontroll, automaatsed meetodid morfoloogia ja süntaksi analüüsimiseks ning tööriistad poolautomaatseks koodide sisestamiseks. (MacWhinney 2000: 15)

### 3.4.2. Koostamispõhimõtted Eestis

Eesti lastekeelekorpuse jaoks pole veel loodud ühtseid koostamispõhimõtteid. Eesti lastekeele korpuse jaoks kogutud andmete töötlemisel ja esitlemisel tuginevad litereerijad CHILDESi eetikanõuetele ja töötamisepõhimõtetele. See tugevdab eesti lastekeelekorpuse litereerijate tööd ning loob usaldusväärse pinnase teaduslike eesmärkide täitmiseks.

### 3.4.3. Küsitluse tulemused

Siinses töös seatud eesmärgi täideviimiseks on kasutatud lisaks korpusuuringule ka küsitlusmeetodit. Küsitlus on leitav töö lisades. Järgneb üldistav kokkuvõte küsitlusest saadud tulemusest.[7]

Eesti lastekeelekorpust on täiendatud ja loodud alates 1994. aastast ning korpuse koostamine jätkub. Koostajate seisukoht on, et ühtsed litereerimispõhimõtted on vajalikud nii korpuse koostajaile kui ka kasutajaile, kelleks on sageli näiteks keeleuurijad. Ühtsed litereerimis- ja transkribeerimispõhimõtted on vajalikud selleks, et analüüsitulemused oleksid usaldusväärsed ja seal sisalduvad andmed võrreldavad.

Küsitluse põhjal püüdsin välja selgitada korpuse koostajate suurimad probleemid korpuse koostamisel. Toodi välja, et raske on otsustada kõnevoorude pikkust, lisaks mainiti mitmeid märgendamisega seonduvaid küsimusi: kuidas lahendada eneseparandused/-täiendused, et need hiljem valeinfot ei annaks; kui palju üldse tuleks erinevaid märgendeid kasutada[8]; kuidas märkida venitusi, kokkuhääldusi, mingisuguse emotsiooniga öeldut (näiteks naerdes, nuttes), kuidas märkida intonatsiooni?

Transkribeerimist mittepuudutavate murekohtadena toodi välja suurt ajakulu – ühe tunni kõnesalvestuse litereerimiseks kulub vähemalt 10 tundi. Samuti osutati sellele, et korpuses võiks olla rohkem erinevaid dialooge[9].

Küsitluses keskendusin ka konkreetsematele juhtumitele: iseseisvate sõnade kokkuhäädus, üneemid, minimaaltagasiside ning kõnevoore puudutav.

Kokku häädatud sõnade puhul on korpuse koostajad lahendanud küsimused sisetundest lähtudes. Kokkuhäädatud sõnad eesti lastekeelekorpuses kirjutatakse lahku või liidetakse +-märgiga. Üneemid ja minimaaltagasiside on uurijad transkribeerinud kuulmise järgi. Mainiti ka seda, et mõned häälitsused on jäetud litereerimata ning et litereerimine võib olla ebatäpne.

---

[7] Küsitluse kokkuvõttes on kasutatud Reili Arguse, Maigi Vija, Sirli Zuppingu ja Helen Kõrgesaare kirjalikke vastuseid.

[8] Esialgu langetasid korpuse koostajad transkribeerimisotsuseid ise. Hiljem, kui saadi MacWhinney käsiraamat CHILDESi korpuse koostamise kohta, selgus, et võimalikke märgendeid on väga palju.

[9] Näiteks soovides uurida hoidjakeelt soost lähtuvalt, on raske leida isasid, kes oleks valmis enda ja lapse vestlust lindistama.

Ebaselgusi märgivad kõik küsitlusele vastanud korpuse koostajad märgiga *XXX*. Kõnevooru märgitakse eesti lastekeelekorpuses kõneleja vahetudes või pausi järgi. Kui paus on tajutav ja pigem pikem, tehakse uus kõnevoor.

Küsitlusest saadud lisainformatsiooni olen taustaks kasutanud ka järgnevas analüüsis, kus olen võrrelnud CHILDESi eesti alamkorpusi.

## 4. CHILDESi eesti lastekeelekorpuse alamkorpuste võrdlus

Oma töös võtsin vaatluse alla kõik CHILDES-i eesti alamkorpused[10], keskendudes neis kirja pandud eri transkriptsioonidele. Analüüsimiseks kogusin lindistuste litereeringutest andmeid seal transkribeeritud tagasisidesõnade, arvude/numbrite märkimise, selgituste, kommentaariridade ja võõrsõnade kohta. Lisaks pöörasin tähelepanu situatsiooni kirjeldustele.

Valisin igast korpusest kolm lindistust: lindistusperioodi algusest, keskelt ja lõpust. Kui ühes korpuses oli lindistatud rohkem kui ühte last, valisin iga lapse kohta ühe lindistuse, millest andmeid koguda (vt tabel 2). Kõik järgnevalt töös esitatud andmed ja näited põhinevad analüüsimaterjali hulka valitud lindistustel.

Tabel 2. **Analüüsiks valitud lindistused iga korpuse kohta.**

| Korpuse nimi | Lindistused |
|---|---|
| Argus | hend01.cha; hend09.cha; hend17.cha |
| Beek | 0_09.24f.cha; 1_01.30m.cha; 2_05.00m.cha |
| Kapanen | 01.cha; 06.cha; 11.cha |
| Kohler | ann170800.cha; car170900.cha; hel240701.cha; hen050900.cha; mar160600.cha; san291201.cha; ste150900.cha; tai180900.cha |
| Kõrgesaar | gregory06.cha; harley11.cha; kaisa01.cha; ruuben03f.cha; andri01.cha; arabella01f.cha; artur01.cha; hellyn01.cha; jaana01.cha; mia01.cha; olivia01.cha; sirlin01.cha |
| Vija | 10724.cha; 20413.cha; 30113.cha |
| Zupping | 1_03.03.cha; 2_02.04.cha; 4_02.13.cha |

[10] Siinses töös kasutatud andmed pärinevad leheküljelt http://childes.psy.cmu.edu/browser/index.php?url=Other/Estonian/ (19.05.2016)

## 4.1. Tagasisidesõnade transkriptsioon eesti lastekeelekorpuses

Eesti alamkorpustes on tagasisidesõnadena kasutatud sõnu *jaa, jah, mhmh, ahah, mkm, jahah, ja-jaa, ekee, jaah, äkää*.

**Tagasisidesõnade transkriptsioon korpuseti**

Arguse alamkorpuses on nii jaatavaid kui ka eitavaid tagasisidesõnu transkribeeritud mitmel moel. Jaatavate sõnade kirjapilt varieerub. Esinevad vormid *ja, jaa, jaah, jahah.* Ka eitavate tagasisidesõnade kirjapilt ei ole alati ühesugune. Eitavaid sõnu on transkribeeritud järgnevalt: *ei, eei.* Neutraalsetest vormidest esineb sõna *ahah.*

Beeki alamkorpuse tagasisidesõnade transkriptsioon on samuti mitmekülgne. Esinevad jaatavad vormid *jaa, jaaah, jaah, mjaah, jahh, jah, mhmm, jaaa, mhmh,* eitavatest tagasisidesõnadest aga *eiiii, eii, eiä.* Sõna *eiä* ütles laps vanuses 1;1.30. Samuti esineb neutraalne tagasisidesõna *ahah.*

Kapaneni alamkorpuses tagasisidesõnade kirjapildist leiab järgmisi vorme: *jah, jaa, mhmh, mkmm, mkm, äkä.* Kasutatud on ka sõna *ahah.*

Kohleri alamkorpuse jaatavad tagasisidesõnad on transkribeeritud kujul *jah, mhmh, jaa, ja.* Eitavatest sõnadest leidub sõna *ei.* Siinses korpuses on neutraalsete tagasisidesõnade varieeruvus suurem: *ahah, okay, ok.*

Kõrgesaare alamkorpuse jaatavate tagasisidesõnade transkriptsiooniviis on mitmekesine: *jaa, jah, mhmh, jajajajaa, jaaaa, jajaaaa, jahh, ähäh, mmhmh.* Eitusüneemidest sõnadest esineb järgmisi vorme: *äkää, mkmm, mkm* ning neutraalsetest tagasisidesõnadest *okei, ahahh, ahah.*

Vija alamkorpuse tagasisidesõnade transkriptsioon on järgmine: *jaa, jah, mhmh, jaah* ning *ahah.* Eitavatest sõnadest *ei.*

Zuppingu alamkorpuses on tagasisidesõnad märgitud järgmiselt: *jah, mhmh, ja-jaa* ning *ekee* ja *ei.* Tagasisidesõnade kokkuvõttev võrdlus on esitatud tabelis 3.

Tabel 3**. Tagasisidesõnade võrdlus korpuste lõikes.**

| Sarnasused | Erinevused |
|---|---|
| Rohkelt jaatavaid tagasisidesõnu: *jaa, jah, mhmh, ahah, mkm, jahah, ja-jaa, ekee, jaah, äkää* | Arguse alamkorpuses ei olnud sõna *mhmh*, mis kõigis teistes oli |
| | Kohleri alamkorpuses transkribeeriti *okay* ja *ok* |
| | Kõrgesaare alamkorpuses *okei* |
| | Zuppingu alamkorpuses ei olnud sõna *ahah*, mis kõigis teistes oli |
| | Beeki alamkorpuses *mjaah* |

**Kokkuvõte tagasisidesõnade transkriptsiooni kohta**

Eesti alamkorpustes on tagasisidesõnu kasutatud võrdlemisi palju. Dialoogipartiklid ongi just suulisele kõnele iseloomulikud (*ahah, mhmh, jah).* Dialoogipartiklid kuuluvad suhtluspartiklite hulka ja nad on omakorda osa pragmaatilistest partiklitest. Dialoogis on need kõik kasutusel selleks, et kuulaja saaks reageerida kõneleja öeldule. (Hennoste 2000b: 50)

Hennoste (2000b: 50) järgi väljendab partikkel *ahah* seda, kui kuulaja jaoks oli jutus midagi uut. *Jah* või *jaa* viitab sellele, et kuulaja nõustub kõnelejaga. Jaatavate tagasisidesõnade esinemine on lastekeele korpuses väga sagedane. Lisaks üksi esinemisele on neid sõnu ka lausete sees ja lõpus. Sageli on ühes kõnevoorus ka mitu jaatavat tagasisidesõna, mis võivad lauses paikneda ka kõrvuti.

Enamasti on neid transkribeeritud *jah, jaa,* kuid esineb ka viise, kus vokaale on järjestikku rohkem kui kaks: *jaaaa, jajaaaa*.

Veel esineb sageli jaatavat tagasisideüneemi *mhmh*. Kasutades partiklit *mhmh,* „annab kuulaja märku, et teine võib edasi rääkida“ (Hennoste 2000b: 50). Eitavaid tagasisidesõnu ja -üneeme esineb eesti lastekeelekorpuses pigem vähem. Eitavaid tagasisideüneeme on transkribeeritud viisil *mkm, äkä* või *ekee.*

## 4.2. Arvude transkriptsioon eesti lastekeelekorpuses

Eesti alamkorpuste litereeringutes (s.t laste ja täiskasvanute kõnes) esineb arve nullist miljonini. Kõige rohkem kasutatakse arve ühest kümneni, vähem suuremaid numbreid.

**Arvude transkriptsioon korpuseti**

Arguse alamkorpuses on arvudest transkribeeritud variandid *üks, kaks, kümme, neli, kuus.* Arve *kümme* ja *neli* on ka laps püüdnud öelda; neid on transkribeeritud järgmiselt: *kumme.*

Beeki alamkorpuses on kasutatud arvude transkriptsioon järgmine: *üks, kaks, kolm, viis, kuus, kaheksa.*

Kapaneni alamkorpuses esineb samade arvude transkribeerimisel ka erinevusi. Näiteks: *üks, üüks.* Ülejäänud arvud olid transkribeeritud ühtemoodi: *kaks, kolm, neli, viis, kuus, seitse, kaheksa, üheksa, kümme, üksteist, kaksteist, kakskümmend.*

Kohleri alamkorpuses on transkribeerimise erinevused täiskasvanu ja lapse kõne vahel. Täiskasvanu öeldud arvud on transkribeeritud alati ühtemoodi: *üks, kaks, kolm, neli, viis, kuus, seitse, kaheksa, üheksa, kümme.* Lapse öeldud numbrid on transkribeeritud järgmiselt: *kak neli kom, kolm kak neli.*

Kõrgesaare alamkorpuses on samu arve transkribeeritud erineval moel. Näiteks kümned on transkribeeritud järgmiselt: *kakskümmend üks null viis, kakskümmend üks ja kolm, üheksakümmend, kakskend kaks kakskend kolm, nelisada kolmkend, kuuskend kaks, kakskend, kolkendviis, kakskend kaks, kolmkümmend kaks, kakskend viis, kakskümmend, kolmkümmend, seitsekümmend, kuuskümmend neli, kolmkümmend kaks, sada kakskümmend.* Teised arvud on transkribeeritud järgmiselt: *seitseteist, kümme tuhat viissada kümme, üksteist, kolmteist, üksteist, kaksteist, kaks, kolm, neli, viis, üks, üheksa, seitse, sada, kaks tuhat neliteist, miljon miljon tosin miljon, kolm, tuhat viissada kümme null*.

Vija alamkorpuses on arvud transkribeeritud ühtemoodi: *üks, kaks, neli, viis, kuus, seitse, kaheksa, üheksa*.

Zuppingu alamkorpuses on arvud transkribeeritud ühtemoodi: *üks, kaks, kolm, üksteist, viisteist*.

**Kokkuvõte arvude transkriptsiooni kohta**

Kõikides eesti alamkorpustes on arvud enamasti transkribeeritud nii, nagu need õiges kirjapildis olema peaksid, s.t ei ole kindel, et need on alati vastavalt häälduse le transkribeeritud.

Kõrgesaare korpuses on transkribeeritud arve ka nii, nagu suulisele keelele kohane: *kolkend viis*[11]*, kakskend, kuuskend.*

Ühe korpuse siseselt on samad arvud transkribeeritud alati samamoodi, välja arvatud Kõrgesaare alamkorpuses, kus kümned on transkribeeritud erinevat moodi.

Kohleri ja Arguse alamkorpustes on lapse öeldud arve transkribeeritud vastavalt nii, nagu lapsed neid on öelnud ning nende järele on lisatud selgitused.

Eesti lastekeelekorpuses transkribeeritakse arve, mis puudutavad protsente, kellaaega, loendamist ning aastaarve. Üldjuhul on sama arvu transkriptsioon alati ühesugune, varieeruvusi esineb vähe. Laste öeldud arvud on transkribeeritud nii, nagu laps neid ütleb. Sellisel juhul on arvu järele lisatud selgitus.

## 4.3. Selgitused eesti lastekeelekorpuses

Eesti alamkorpustes kasutatakse rohkelt selgitusi. Selgitused on nagu tõlked, mille abil avatakse n-ö arusaadav sõnavorm või keeleline väljend. Selgituse kirjutab transkribeerija kandilistesse sulgudesse, lisades sõna või sõnade ette sulgude sisse kooloni või võrdusmärgi. CHILDESi manuaali (MacWhinney 2000: 194) järgi tuleb kandilistesse sulgudesse lisada koolon, kui soovitakse asendada ning võrdusmärk siis, kui on vaja selgitada.

Selgitusi lisatakse näiteks nende sõnade juurde, mis on ortograafilisest vormist teistmoodi öeldud – siis lisatakse selgitussulgudesse sõna ortograafiliselt õige vorm. Ka lapsekeelsete sõnade järele lisatakse selgitusi, näiteks kui laps viitab koerale sõnaga *aua,* lisatakse selgitussulgudesse *koer*.

[11] Arvule *kolkend viis* on järele lisatud selgitus: *OBS: kolkendviis [: kolmkümmend viis]. Teistele suulisele keelele kohaselt hääldatud ning transkribeeritud vormidele selgitusi järele lisatud ei ole.

**Ülevaade korpuseti**

Arguse alamkorpuses kasutatakse täiskasvanute keelest erinevate vormide puhul ainult koolonit. Näiteks Arguse alamkorpuses on selgitatud need osad, kus laps ütleb *s*-tähe asemel *t*-tähe. Tõenäoliselt ei oska laps selles vanuses veel õigesti hääldada ja seetõttu kipuvad need sageli vahetusse minema.

Näiteks **CHI: tisse [: sisse] tuli; *CHI: teda [: seda]; *CHI: tiin [: siin] pime oo [: öö], emme; *CHI: mäu taba [: saba]; *CHI: uhh [: uks] tinna [: seal]; *CHI: ästi tüga [: sügav]; *CHI: tiis [: siis] enna tudi*.

Selgitusi on lisatud ka siis, kui on ära jäetud mõni häälik sõna algusest või sõna lõpust: **CHI: alla tule [: tulen]; *CHI: aia [: aias]; *CHI: mina , mina kooli [!] akka [: hakkan minema]; *CHI: enna [: venna] oma*.

Lapsekeelsete sõnade selgitused Arguse alamkorpuses on järgmised: **CHI: eia aa [: ei saa]; või viu viu autot [: kiirabi]; *EMA: mäu [: kass] on seal või*. Viimase näite puhul on selgitus lisatud ema kõnesse, kuid kui laps ütleb *mäu*, siis seda enam selgitatud ei ole: **CHI: Ninnu tätte [: kätte] mäu kaa.*

Beeki alamkorpuses ei ole selgitusfunktsiooni üldse kasutatud.

Kapaneni alamkorpuses on selgituse transkribeerimisel kasutatud ainult võrdusmärki. Selgitused on üldjuhul lapse kõnelemise juures, kuid esineb ka seda, et ema öeldut on selgitatud. Näited lapse kõne juurde lisatud selgitustest: **CHI: eme [= emme]; *CHI: kii [= diktofon]; *CHI: linni linni [= lindistab]; *CHI: dia tah [= diktofon] dika [= diktofon]; *CHI: maa sinu eest usta [= tõsta]; *CHI: ta nääb me tustamee [= tõstame], mis on tomatigaa; *CHI: tsin [= siin] pildi peal; *CHI: mina lähen toobikka [= toob ikka], hobust lähen toob emmile [= emmele], län [= lähen] toon emmile [= emmele] väikest obust; *CHI: aa iisuälatavad [= isuäratavad] maasikad ka, xxx oopis oopis polgandit; *CHI: kassa leidsid juba mõned munaestid [= munarestid]; *CHI: kas sa jõõd [= jõuad] seljakotti kanda; *CHI: emme ma panen selle kaanikausi [= kraanikaussi] sest see on vaja äla pista [= pesta] emme.*

Mõningatel juhtudel on vaja selgitus lisada ka vanema kõne juurde. Näiteks kui kasutatakse lapsekeelseid vorme: **MOT: oo pall [= õhupall] lendab*. Vanemate kõne juurde lisatakse selgitusi ka siis, kui nende suuline keel erineb kirjakeelest, näiteks on kuulamise järgi jäänud ära mõni täht sõnast või räägitakse kõnekeelele omaselt

lühikeselt ja ei hääldata kõiki sõnu välja: **MOT: mmpanen [= ma panen] su siia hop; *MOT: votoka [= fotoka] panid talle peale või; *MOT: no kusta [= kus ta] on*.

Kohler kasutab oma alamkorpuses pigem koolonit, mõnel üksikul korral ka võrdusmärki. Selles alamkorpuses on tihti märgitud ka lapse ja täiskasvanu üksteise imiteerimised. Selgitused on lisatud lapse täiskasvanupärasest keelest teistmoodi öeldud sõnadele: **CHI: kolk [: kork]; *CHI: ei ane [: pane]; *CHI: jah (.) pogand [: porgand] on; *CHI: suul [: suur] auto; *CHI: naks+naks [= käärid] vetab [: võtab]; *CHI: palnda [: paranda] äla [: ära]; *CHI: olut [: õlut]; *CHI: kak [: kaks] neli kom [: kolm]*; **CHI: lipatiinud [: lepatriinud]; *CHI: üleva [: üleval]; *CHI: õigepidi [: õigetpidi]*.

Selgitusi on lisatud ka siis, kui täiskasvanu räägib lapsekeelseid sõnu: **KAJ: äkki tahad lutut [= lutipudelit]*.

Kõrgesaare alamkorpuses on kasutatud nii koolonit kui ka võrdusmärki. Täiskasvanu kõnes on parandatud kõnekeelseid vorme, näiteks kui sõna algusest on ära jäetud *h* või kui sõna pole terviklikult välja öeldud: **OBS: noh mis sa joonistama akkad [: hakkad] mulle; *OBS: a [: aga] kas te selle koduse töö kontrollisite ära või; *OBS: kõlab teistmoodi oopis [: hoopis]; *CHI: teeks iljem [: hiljem]; *OBS: mis sa teed seal esimeses klassis igav akkab [: hakkab] ju sedasi; *OBS: mm a [: aga] selle kõige parema jätsid ikka enda teada jah.*

Ka lastekõnes esineb kõnekeelseid lühendatud vorme või sõna alguse *h* ärajätmisest: **CHI: ei ei ma vaatasin laua alla ja vaatasin enda ümbrusse ei old [: olnud] vaatasin ka laua peale kuskile alla ei old [: olnud] üldse; *CHI: siuke [= selline] lugu oli et ee sõna nutma et üks õpetaja et üks õpetaja viis lapsed muuseumisse ja siis lapsed ei saand aru mis sõna see on; *CHI: siuke [= selline] värk siuke [= selline] värk oli seal et seal oli kirjas kuul aastal kaks tuhat neliteist ja seal räägiti võru keeles plaadi peal jaa jah ja räägiti ka ja räägiti võru keeles ka õpikus; *CHI: ma teen ästi [: hästi] kiirluubis või noh nimodi kiirluubis ästi [: hästi] kiirelt.*

Selgitustesse on lisatud ka laste täiskasvanupärasest keelekasutusest teistmoodi öeldud sõnade parandused: **CHI: laamit [= raamat] tulid ka välja; *CHI: ei see on see on ka keegi [= kellegi] kodu; *CHI: ei suju [= suru]; *CHI: kommapade [:*

*kommipaber]; *CHI: viskan see kommipade [: selle kommipaberi]; *CHI: pügipassi [: prügikasti]*.

Vija alamkorpuses on selgituste kirjapanekul kasutatud pigem võrdusmärki, kooloni kasutamist esineb vähem. Lapsekeelsete vormide järele on lisatud selgitused: **CHI: kiss+kiss [= kass]; *CHI: kuts [= koer] kuts [= koer]; *CHI: patt [: part]; *CHI: jäku [: jänku]; *CHI: ninni [= nina]*.

Ka on lisatud selgitused, kui nimesid öeldakse tegelikust teistmoodi: **CHI: Eki [: Eiki]; *MOT: Antsu [= Andreas]; *CHI: Elle [= Erel]; *CHI: Kist [= Kristjan]*.

Selgitustega on parandatud lapse öeldud sõnu, mis erinevad täiskasvanupärastest vormidest: **CHI: Antsu [= Andrease] tahab teist piilatsi [= pliiatsit] nüüd; *CHI: kas sina tahad tardrikust [: taldrikust] süüa või tassist; *CHI: kipsi [= küpsis]; *CHI: puppu [= nuppu]*.

Zuppingu alamkorpuse selgitused on transkribeeritud võrdusmärkidega. Zuppingu korpuses ütles laps *r*-tähe asemel *l*-tähte, nende sõnade järele olid lisatud selgitused: **CHI: suulemat [= suuremaid] pannkooke; *CHI: kas me võime ükskold [= ükskord] ploovida [= proovida] neid minna vaatama; *CHI: neid neid vesilattaid [= vesirattaid]; *CHI: nende vesilataste [= vesirataste] seest; *CHI: tegin seal mällaks [= märjaks] kus onu oli; *CHI: vaata tegin siit äla [= ära]*.

Täiskasvanu lapsekeelsed sõnad on ära selgitatud: **MOT: emme tegi atsihh [= aevastas]; *MOT: päike, mäletad, päike, ja see on notsu põssa põrsas siga, nii palju nimesid, kana kaa-kaa, kanade käest saame mune, vaata, ja siis munad lööme katki koks ja saame teha omletti, täna sa sõid omletti hommikul, nämm-nämm [= maitsev], oli nämm-nämm [= maitsev]; *MOT: jah ole minu opsas [= süles] ole minu süles, kui sa keerad ennast ära siis on emmel raske hoida, vaat nüüd tuleb küll mingi asi, mis see oli, sõitis mööda midagi, kas sa ei pannud tähele; *MOT: no anna kätu [= käsi] mulle.*

Lapse püüded öelda sõnu, mis tal veel välja ei tule: **CHI: takta [= diktofon] takta [= diktofon]; *CHI: täät [= tädi] täkto [= diktofon] ää; *CHI: pepp [= hobune] pepp [= hobune]; *CHI: amm [= armas]; *CHI: lööh [= lammas]*.

Täiskasvanu kõnekeelsed vormid: **MOT: mis kell nad siis lähvad [= lähevad]; *MOT: tegelt [= tegelikult] nad ei kuule.* Selgitusridade kokkuvõttev võrdlus on esitatud tabelis 4.

Tabel 4. **Selgitusridade võrdlus korpuste lõikes.**

| Sarnasused | Erinevused |
|---|---|
| Kõikides alamkorpustes, välja arvatud Beeki alamkorpuses, on kasutatud selgituste funktsiooni | Beeki alamkorpuses ei ole litereerija selgituste funktsiooni üldse kasutanud |
| | Arguse alamkorpuses on litereerija kasutanud selgituste lisamiseks ainult koolonit |
| | Kapaneni ja Zuppingu alamkorpustes on selgituste lisamiseks kasutatud ainult võrdusmärki |
| | Nii koolonit kui ka võrdusmärki on kasutatud Kohleri, Kõrgesaare ja Vija alamkorpustes |

**Kokkuvõte selgitustest**

Selgitused on kirja pandud kõikides eesti alamkorpustes, välja arvatud Beeki alamkorpuses. Selgituste lisamiseks on kasutatud nii võrdusmärki kui ka koolonit.

Kohleri alamkorpuses paistab valitud andmete põhjal, et koolonit on kasutatud asenduseks. Lapse täiskasvanupärasest keelekasutusest erinevalt öeldud vormidele on lisatud järele selgitus, kasutades koolonit. Võrdusmärki on kasutatud aga neil puhkudel, kui täiskasvanu kasutab lapsekeelseid sõnu.

Kõrgesaare alamkorpuse andmete põhjal ei ole võimalik üheselt aru saada, millal litererija on kasutanud koolonit, millal võrdusmärki. Transkriptsioonimärke kasutatakse nii kõnekeelsete vormide kui ka täiskasvanupärastest vormidest hälbimise puhul.

Vija alamkorpuse puhul tundub olevat järgitud põhimõtet, et selgituste lisamiseks kasutatakse üldjuhul koolonit. Võrdusmärki kasutatakse siis, kui laps püüab öelda sõnu, aga need sisaldavad pisivigu.

Selgitused on pandud nii täiskasvanu kui ka lapse kõne järele. Täiskasvanu kõne juurde on selgitused pandud näiteks siis, kui täiskasvanu räägib lapsekeele sõnadega või kõnekeelselt.

Laste kõnevoorudes on selgitusi rohkem, mis on ka ootuspärane, sest laste kõne on arenemises. Selgitused on lisatud püüdlustele öelda nimesid või sõnu, kui need on öeldud kergete eksimustega või täiskasvanupärasest vormist oluliselt teistmoodi.

## 4.4. Kommentaariread eesti lastekeelekorpuses

Kommentaariread lisatakse transkribeeritud dialoogide sisse. Dialooge lugedes on kommentaari ees märgend *%com*. Kommentaariread tähendavad uurija poolt lisatud kommentaare, mis annavad infot dialoogi või lindistusruumis toimuva kohta. Kommentaarid täidavad erinevaid eesmärke.

Need annavad täpsustusi dialoogides mainitud informatsiooni kohta, kirjeldavad parasjagu toimuvat tegevust või hääli, võivad viidata vestluses osalejate emotsioonidele või füsioloogilistele häälitsustele. Litereerija otsustab, millal on kommentaariridade lisamine vajalik ning millist eesmärki see parasjagu täidab.

Reili Argus on kirjutanud, et lastekeele, nagu suulise keele puhul samuti, ei ole ainuüksi verbaalse info põhjal võimalik mõista, millest konkreetsel hetkel räägitakse, ja just sellepärast tuleks kasutada kommentaariridа (Argus 2007: 68).

**Kommentaariridade transkriptsioon korpuseti**

Arguse alamkorpuses on kommentaariridu kasutatud mitmesuguste täpsustuste tegemiseks. Üldjuhul kasutatakse kommentaariridu siis, kui dialoogi põhjal ei ole võimalik päris täpselt aru saada, mis parasjagu toimub. Näiteks on kommentaariridades selgitatud toimuvat tegevust: *%com: paneb paberid sahtlisse; %com: ema imiteerib last; HEN läks teise tuppa autosid tooma; HEN poeb kappi ja mängib kapiustega; HEN jookseb magamistuppa; %com: magamistoas on tekk , millel on kassi pilt; %com: HEN tahab teki pildi pealt kassi sülle võtta.*

Kirjeldatud on emotsioone ja hääli või häälitsusi: *%com: kiunub; %com: Hendrik kiunub; %com: Hendrik nutab; %com: teeb politseiauto häält; %com: HEN naerab; %com: MAR naerab; EMA naerab; HEN sõidab autoga ja põriseb; HEN kilkab; %com: HEN teeb mingeid hääli*.

Samuti on Arguse alamkorpuse kommentaariridades kirjeldatud vestlusse puutuva tausta: *%com: nähtavast HEN ei mäleta enam haiglasolekut*; *%com: ema pani eelmisel õhtul Hendrikule kogemata vanema venna pidzaamapluusi selga; %com: pidzaamapluusil on lennukite ja autode pildid.*

Beeki alamkorpuses on kommentaariridu kasutatud olukordade ja tegevuste kirjelduseks: *%com: CHI tahab mobiili kotti suhu panna; %com: FAT annab CHI-le mänguasja; %com CHI taob mänguasja vastu maad; %com FAT hakkab CHI riidesse panema; %com: CHI mängib telefonikotiga; %com: MOT ja CHI korjavad maast mänguasju; %com: CHI vaatab enda peegeldust plekist karbi kaanelt.*

Beeki alamkorpuses on kommentaariridades kirjeldatud ka hääli ja emotsioone, sealhulgas füsioloogilisi helisid: *%com: CHI ja FAT matsutavad; %com: CHI undab järelejätmatult; %com: CHI nutab; %com: CHI aevastab; %com: CHI nutab FAT annab CHI-le putru; %com: CHI teeb musi häält; %com: koer ägiseb läbi une; %com: FAT ja CHI naeravad; %com: CHI läkastab.*

Kapaneni alamkorpuses kirjeldatakse kommentaariridadel toimuvaid tegevusi: *%com: otsivad pabereid ja joonistusvahendeid; %com: CHI ja MOT jagavad pliiatseid, kes missugusega hakkab joonistama; %com: CHI jalutab nukku mööda lauda; %com: ema toimetab; %com: CHI kõnnib vahepeal; %com: ema kõnnib lapsest eemale; %com: ema joob ja tuleb siis lapse juurde tagasi.*

Kommentaariridadega kirjeldatakse inimeste hääli ja teisi kõrvalisi helisid, mida ei tee inimesed. Näited inimeste häältest ja häälitsustest: *%com: Martina köhib; %com: CHI häälitseb midagi; %com: CHI haigutab; %com: laps nohiseb; %com: ema köhib; %com: laps matkib ema köhimist*. Lisaks on märgitud hääli, mis tulevad ümbrusest: *%com: kostub pliiatsi sahinat; %com: kostub kõva ajalehekrabinat; %com: kostub truki krõps; %com: ema ja Martina räägivad midagi vaikselt omavahel, on kuulda nõudega kolistamist, veepahinat ja kohviaparaadi häält; %com: kostub vee pahinat, ema peseb midagi; %com: Martina ütleb midagi, kuid läbi kohvimasina puhastamise surina ei ole kuulda; %com: kostub üsna kõva kohvimasinast tulevat lurinat.*

Ära on märgitud ka kõnelejate omavahelised vaheleräkimised, katkestused ning hetked, mil lindistaja ei kuule või ei saa aru, mida öeldakse: *%com: CHI ütleb midagi, ei ole aru saada; %com: laps ütleb lausungi lõppu veel midagi, ei ole aru saada;*

*%com: FATi lausungi lõppu ei olnud kuulda; %com: ema lause jääb pooleli, isa ütleb midagi; %com: isa ütleb midagi, ei ole kuulda; %com: laps räägib midagi sosinal, ei kuule; %com: vahepeal keegi ei räägi; %com: keegi ei räägi vahepeal; %com: ema segab vahele.*

Kapaneni alamkorpuses on kommentaarides märgitud ka täpsustused, kuidas keegi midagi ütleb: *%com: CHI räägib lausungi lõppu peenikese häälega kiiresti silpe; %com: laps ütleb sõnad järjestikku üsna kiiresti; %com: CHI proovib kolme eri viisi öelda sõna õun; %com: Martina ütleb nii sõna alguses kui ka keskel hästi pehme r-hääliku.*

Kohleri alamkorpuses on kommentaariridu võrreldes teiste alamkorpustega oluliselt vähem. Selles alamkorpuses on lindistatud 8 last, mõne lapse lindistuses ei esine mitte ühtegi kommentaaririda.

Kohleri alamkorpuse kommentaariridades on kirjeldatud toimuvaid tegevusi: *%com: vaadatakse albumit; %com: MOT ja HEL laulavad; %com: kõik lähevad kööki ja MOT teeb juua.*

Selles korpuses esines ka täpsustusi diktofoni puudutavate asjaolude kohta: *%com: paus lindistuses, sest vahepeal tuleb maki sisse uued patareid panna; %com: diktofon seiskub, kui keegi ei räägi.*

Samuti on ära märgitud mõned hääled: *%com: keegi krõbistab akna taga; %com: sõrm jääb auto vahele ja TAI hakkab nutma; %com: nõudepesumasin suriseb.*

Lisaks mainitakse kommentaariridadel ära, kui lisandub inimesi: *%com: tuleb Retti; %com: saabub CAR; %com: tuleb VAL.*

Kõrgesaare alamkorpuses kirjeldatakse kommentaariridadel parasjagu toimuvaid tegevusi: *%com: CHI jookseb raamatu järele; %com: FAT teeb köögilaua peale joonistamiseks ruumi; %com: MOT hakkab ise laulma, aga CHI tegeleb edasi rongiga; %com: FAT segab vahepeal juurvilju pannil; %com: CHI otsib pilti.*

Kommentaariridadel täpsustatakse vestluses mainitud inimesi või fiktiivseid tegelasi: *%com: Ralf on pinginaaber; %com: tegemist on mereröövlikapteniga filmist Kariibi mere piraadid; %com: Mikk on vanema venna klassivend; %com: Kät on lindistaja, kellel külas ollakse; %com: Mona on kohalik koer.*

Lisaks annavad kommentaariread ülevaate lindistatud inimeste häältest ja emotsioonidest: *%com: CHI hakkab tantsima nii suure hooga , et on maha kukkumas , MOT naerab; %com: CHI hakab karjuma ja OBS võtab ta sülle; %com: CHI itsitab; %com: CHI teeb pahurdamishäälitsusi; %com: CHI naerab kõlava häälega; %com: naeravad koos.*

Siinses alamkorpuses on märgitud ära ka see, mis jääb litereerimata. Kommentaariridadele on kirjutatud, et ei litereeri: *%com: telefon heliseb ja MOT läheb elutuppa vastama. ei litereeri; %com: CHI tsiteerib vanema õe kirjutatud luuletuse algust. Ei saa aru ja ei litereeri; %com: MOT räägib eemal telefoniga, ei litereeri; %com: viimane oli mõeldud koerale. MOT püüab selgitada vanemale tütrele, miks too ei tohi kööki tulla ja tekib väike sõnelus, ei litereeri; %com: vanem tütar tikub vägisi segama, MOT läheb selgitustööd tegema, ei litereeri.*

Kõrgesaare alamkorpuses märgitakse kommentaariridadel ära ka need kohad, kui litereerija ei ole aru saanud, mida öeldakse: *%com: CHI jutust ei saa aru; %com: MOT-i teksti algus mattub nugade-kahvlite-lusikate kolina alla; %com: OBS läheb eemale ja räägib midagi , mida pole kuulda; %com: CHI räägib nii sogaselt, et ei ole täpselt aru saada, mida ta ütleb.*

Vija alamkorpuses valitud lindistustes kommentaariridu ei esinenud.

Zuppingu alamkorpuses on märgitud hääled ja emotsioonid: *%com: MOT aevastab; %com: PAU tuleb, põrand müdiseb; %com: CHI ja MOT teevad musitamise hääli; %com: CHI teeb auto häält; %com: kärbes sumiseb; %com: CHI rõõmustab; %com: CHI aevastab.*

Märgitud on ka, kui osa lindistatud vestlusest jääb litereerimata: *%com: MOT ja PAU räägivad isiklikke teemasid, ei litereeri; %com: täiskasvanute omavaheline vestlus (ei litereeri).*

Kirjeldatud on tegevusi, mis lindistamise ajal toimuvad: *%com: MOT ja CHI läksid kööki kohvi tooma; %com: PAU puhub CHI varba peale ja mürab temaga; %com: CHI serveerib emale mängult toitu, toit kukub kandikult maha; %com: CHI kõlistab klaasist päevalille vastu aknaklaasi; %com: CHI seletab isale, mida ta hetkel näeb ja mida aknast nägi.*

Kui kõnest ei ole täpselt aru saada või ei ole seda kuulda, on see kommentaariridadel välja toodud: *%com: CHI läheb mööda silda kaugemale ja räägib midagi; %com: CHi jätkab sosinal; %com: CHI räägib midagi.*

Kommentaariridade korpustevahelised sarnasused ja erinevused on kokkuvõtvalt välja toodud tabelis 5.

Tabel 5. **Kommentaariridade võrdlus korpuste lõikes.**

| Sarnasused | Erinevused |
|---|---|
| Eesti alamkorpustes kasutatakse kommentaariridu | Vija alamkorpuse juhuslikult valitud lindistustes ei ole kommentaariridu kasutatud[12] |
| Kommentaariridu kasutatakse, et märkida häälitsusi, hääli, emotsioone, olukorra või tausta kirjeldamiseks | Zuppingu ja Kõrgesaare alamkorpustes on märgitud, kui ei litereeri |
| Alamkorpustes märgitakse ära, kui kõnest ei saa aru | |

**Kokkuvõte kommentaariridade transkriptsiooni kohta**

Eesti alamkorpustes on kommentaariridu kasutatud aktiivselt. Need on täpsustavad ja selgitavad ning annavad ülevaate, mida lindistuse ajal parasjagu tehakse.

CHILDESi manuaali (MacWhinney 2000: 83) järgi ei ole kommentaarireal ühte konkreetset eesmärki, sellel on üldine funktsioon. Ta täpsustab ka, et pigem tuleks kommentaariridadel kasutada koodide asemel sõnu, et programm seal sisalduvat vigadeks ei märgiks (MacWhinney 2000: 83).

Kommentaariread aitavad luua ka ülevaadet, kui midagi on litereerimata jäetud või kui litereerija pole kõnest aru saanud.

[12] Siinses uurimuses vaatluse alla võetud lindistustes kommentaariridu ei esinenud, teistes Vija korpuse lindistustes kommentaariridu siiski esineb.

## 4.5. Võõrsõnade transkriptsioon eesti lastekeelekorpuses

Järgnevalt antakse ülevaade eesti alamkorpustes esinevate võõrsõnade transkriptsioonidest. Vaatluse alla on võetud need sõnad, mis algavad häälikutega *g, b, d* ning mõningad sõnad, mis sisaldavad tähte *š* või *ž, f.*

**Võõrsõnade transkriptsioon korpuseti**

Arguse alamkorpuses esineb võõrsõnadest üks võõrsõna: *g*-ga algav *garaaž*, mis on transkribeeritud järgmiselt: **CHI: ussi laasi [: garaazi] pane ; *EMA: paned garaazi.*

Beeki alamkorpuses on samuti sõna *garaaž*, mis on transkribeeritud nii: **MOT: teise viis garaasi*.

Kapaneni alamkorpuses esinevad võõrsõnadest *diktofon, diivan, film, buss, tšilli, banaan, beebi*. Nende sõnade transkribeerimisel on tõenäoliselt järgitud ortograafiat, mitte häädlust. Järgnevad näited transkribeerimisest: **MOT: jaa see diktofon teeb oma tööd; *MOT: film; *MOT: diivani peal; *FAT: banaan on vanaks läind jah; *CHI: seda et on vaja beebi üles älatada; *FAT: ei ära tšillit pane; *MOT: venna läks bussiga.*

Lisaks esinesid mõned võõrkeelsed nimed, mis on transkribeeritud siinses alamkorpuses vastavalt häädlusele. Näiteks *Donald Duck* ning *iPod*: **MOT: oi mis sa panid vennal siin ühe filmi käima, hmm sa oled väga nutikas, vaata siin on nüüd üks film, näedsa sa panid aipoodis filmi käima, emme ei olegi vaadanud, mis filmid seal on, tasakesi; *CHI: olgu las nad siis pikutavad siin diivanil Minniga koos; *CHI: Ma Donald Takki.*

Kohleri alamkorpuses on võõrsõnadest kasutatud kaks sõna: *bensiin* ja *diivan*, mis on transkribeeritud nii nagu kirjakeeleski: **KAJ: a' miks ta seisab (.) bensiin sai otsa või; *KAJ: diivani alla (.) võtad välja või.*

Kõrgesaare alamkorpuses on transkribeeritud sõnad *oranž, firma, fantaasia, diivan, doomino, bemoll, grillvorst, banaan, batuut.* Siinses alamkorpuses ei ole korpusesiseselt kõiki võõrsõnu alati samal viisil transkribeeritud. Erinevaid variante esineb sõna *oranž* ning *diivan* puhul: **FAT: mõned mustad täpid ja natukene oraantši; *CHI: nii jah oranši mm ma ei tea teeme nii et sina joonistad näiteks joonistame jahi kuule; *MOT: diivan; *FAT: kuule tibu mine pane need oopis tiivani peale sinna tiivani peale pane*

*sokid vaata seal on tiivan nääd jah tubli tüdruk tiivani peale sokid; *FAT: issi maub ikka vaata issi istub diivani peal issi mahub siia.*

Teised võõrsõnad, mis alamkorpuses esinesid, on transkribeeritud korpusesiseselt ühtmoodi või esineb neid vaid ühel korral: **FAT: on firmadel on jahid ja on eraisikutel on jahid; *FAT: jaa ja sa saadki nööd akata seda nüüd täitma oma värvidega kui sa tahad oma fantaasia järgi; *CHI: diees on märk mis kõrgendab nooti vist või mis asi on bemoll; *CHI: millest on valmistatud pikkpoiss kas aa suitsulihast bee grillvorstist või tsee hakklihast; *EMA: natuke läheb aga pea-aegu sa tead vaata kus see banaani täht käib; *CHI: sin väikse batuudi ka oli.*

Siinses alamkorpuses on transkribeeritud ingliskeelseid sõnu ja nimesid häälduse järgi. Näiteks *dark black, marine blue, national flags, New Zealand, Volkswagen, Peugeot* ja *Mitsubishi Evo* on transkribeeritud nii: **CHI: dark bläkk; *FAT: see on mariin bluu; *FAT: okei mul on üks mäng näššönal fläägs; *FAT: nii njuu ziiland; *FAT: see oli Volksvaagen aah tuleb veel autooo see oli Pežoo; *FAT: see oli Mitsubiši Evo see on päris äge auto.*

Vija alamkorpuses esinevad võõrsõnadest sõnad *diktofon, banaan, buss* ja *bassein*, mis on transkribeeritud vastavalt kirjakeelele: **MOT: see on diktofon , ära sa puutu seda , eksju; *MOT: banaanid on jah; *CHI: bussi; *MOT: käisid basseinis , mis sa seal tegid.*

Zuppingu alamkorpuses on võõrsõnad *diktofon, diivan, garderoob, ketšup, banaaniplombiir* ja *draakon*. Nimetatud võõrsõnad on transkribeeritud järgmiselt: **MOT: diktofon mitte taat, see on diktofon, diktofon; *MOT: pähe peab selle panema jah, mhmh see kiiver on nii raske, sellega sa kukud maha, kukud diivani pealt maha; *MOT: see on väga raske kiiver, las ta olla siin, otsime sulle oma kiivri, meil on garderoobis veel kiivreid, vaatame; MOT: paned ketšupit peale; *MOT: näe siin on eriti head siin on need banaaniplombiir ja Vanilla Ninja mis sulle väga meeldis; *CHI: ee kas see daakoni [= draakoni] või; *MOT: ei mitte draakoni vaata seda see sulle maitses.* Võõrsõnade transkriptsiooni võrdlus on esitatud tabelis 6.

Tabel 6. **Võõrsõnade transkriptsiooni võrdlus korpuste lõikes.**

| Sarnasused | Erinevused |
|---|---|
| Ühe alamkorpuse piires on sama sõna transkribeeritud ühtmoodi, välja arvatud Kõrgesaare alamkorpuses | Kõrgesaare alamkorpuses on ühte ja sama sõna transkribeeritud erineval moel, näiteks *diivan* ja *tiivan* |
| Võõrsõnad on enamasti transkribeeritud vastavalt sõna ortograafiale | Kapaneni ja Kõrgesaare alamkorpuses on võõrkeelseid nimesid või sõnu transkribeeritud vastavalt häälduse (*iPod* vs *aipood*). Beeki korpuses on sõna *garaaž* transkribeeritud *garaas* |

**Kokkuvõte eesti alamkorpuste võõrsõnadest**

Siinse töö analüüsiks valitud materjali hulgas ei ole kasutatavate võõrsõnade hulk kuigi suur, kuid need sõnad, mida kasutatakse, ühtivad erinevates korpustes. Näiteks väga sageli tuli ette sõna *banaan* ja *diivan.*

Üldjuhul transkribeeritakse sõnad nii, nagu neid kirjutatakse, kuid mõnel juhul esineb seda, et transkribeeritakse häälduse järgi. Näiteks *diivani* asemel *tiivan* ning võõrkeelsed sõnad mitte originaalpildis vaid samuti häälduse järgi.

Mõneti esineb ka tendentsi, et korpusesiseselt transkribeeritakse sama sõna erineval moel.

# 5. Järeldused

Eesti lastekeelekorpuses esineb transkribeerimisel korpustevahelisi erinevusi. Korpuse praegused koostajad on üksmeelselt arvamusel, et kokku lepitud transkribeerimispõhimõtted on väga olulised ning vajatakse eesti keelele kohast litereerimisjuhendit. Siinses töös võeti transkribeerimise erinevusi silmas pidades vaatluse alla tagasisidesõnad, arvud, selgitused, kommentaariread ja võõrsõnad.

Töö tulemusel selgus, et tagasisidesõnade transkriptsioon erineb näiteks vokaalide kirjapaneku pikkuse poolest (*jaa*/*jaaaa*) või sõnakasutuse poolest (*mkm, äkä, ekee, ei*). Tartu Ülikooli suulise kõne korpuse kodulehel on välja toodud, kuidas transkribeerida tagasisidesõnu: *jaa, mhmh, mqm.* Lastekeele kõnesalvestuste transkribeerimisel võiks edaspidi kokku leppida, kuidas süsteemselt kirja panna eri tagasisisesõnu.

Arvud on üldjuhul transkribeeritud nii, nagu õige kirjapilt ette näeb. Pole kindel, kas need on alati transkribeeritud vastavalt häädusele. TÜ suulise kõne korpuse kodulehel antud soovituse kohaselt tuleks arvud kirjutada sõnadega välja häälduse järgi: *kakskend=kaks, kaeksada=kolgend=öeksa.* Lastekeelekorpuses peab sellise märkimisviisi järel esitama ka selgituse ehk kirjakeelse vormi.

Selgitused on eesti lastekeelekorpuses märgitud ebaühtlaselt. Alamkorpustes on selgituste märkimiseks kasutatud nii koolonit kui ka võrdusmärki. Ei ole võimalik üheselt aru saada, millistel juhtudel missugune märk kehtib. CHILDESi manuaali (MacWhinney 2000: 194) järgi tuleb kandilistesse sulgudesse lisada koolon, kui soovitakse asendada ning võrdusmärk siis, kui on vaja selgitada.

Kommentaariridu on eesti alamkorpuses kasutatud rohkelt. MacWhinney (2000: 83) järgi täidab kommentaaririda üldist funktsiooni. Eesti lastekeelekorpuse kommentaariread on täpsustavad, selgitavad ja ülevaatlikud.

Võõrsõnade puhul esineb tendentsi, et ühes korpuses transkribeeritakse üks sõna erineval viisil. Transkribeeritud on sõnu nii ortograafiapäraselt kui ka häälduspäraselt. TÜ suulise kõne korpuse kodulehel on kirjas, et *g, b, d* võõrsõnade alguses tuleks kirjutada kirjapildile vastavalt: *g, b* ja *d* abil.

Lisaks täheldasin andmetega töötamisel, et onomatopoeetilisi häälitsusi või silpe ning korduvaid sõnu on märgitud nii sidekriipsu kui ka tühikuga. Litereerijad on arusaamatud kohad tähistanud märgendiga *XXX*.

Töö tulemusel selgus, et eelkõige esineb erinevusi tagasisidesõnade, arvude ja võõrsõnade transkribeerimisel. Erinevusi ilmneb nii ühe korpuse sees kui ka korpuste vahel. Eesti lastekeelekorpuse koostamisel tuleks transkribeerimisel silmas pidada esmalt seda, et korpusesiseselt oleks ühed ja samad sõnad/väljendid jmt transkribeeritud ühtemoodi. See välistaks korpusesisese varieerumise. Tagasisidesõnade, arvude ja võõrsõnade transkribeerimisel võiks lähtuda suulise kõne korpuse kodulehelt leitavast transkribeerimisjuhendist.

Praegu on selgituste puhul probleem märkide kasutamise erinevuses. Edasisel korpuste koostamisel tuleks silmas pidada MacWhinney (2000: 194) juhendis seisvat: asendamisel kasutada koolonit ning selgitades võrdusmärki.

## Kokkuvõte

Andmepanka CHILDES võib pidada usaldusväärseks andmekoguks. Kõik otsused, mida korpuse koostaja andmete sisestamisel ja transkribeerimisel teeb, mõjutavad analüüsi tulemusi. Seega on väga oluline, et litereerimisel lähtutaks ühistest ja põhjendatud kokkulepetest.

Siinse bakalaureusetöö üldine eesmärk oli kaardistada CHILDESi eesti lastekeelekorpuse hetkeseis ning alleesmärk uurimise käigus välja selgitada, millised erinevused on alamkorpustes rakendatud litereerimispõhimõtetel ja millest need erinevused tulenevad. Lisaks anti uurimistulemuste põhjal soovitusi edasiseks tööks CHILDESi eesti alamkorpuste koostamisel.

Siinses uurimuses kasutati kvalitatiivset lähenemist. Analüüsiti korpuse materjali ning küsitluse põhjal selgitati välja litereerimisega seotud probleemid korpusekoostajatel. Empiirilises osas kasutati andmetena CHILDESi eesti alamkorpuste litereeringuid ja korpusekoostajate küsitlusi.

Korpusest kogutud andmete analüüsi tulemusel järeldus, et ühistel alustel vaatluse alla võetud faktorite transkriptsioon erines korpuseti. Võrdlusel vaadeldi tagasisidesõnade, arvude, selgituste, kommentaariridade ja võõrsõnade transkriptsiooni.

Varem ei ole CHILDESi eesti lastekeelekorpuse litereeringuid ja nende loomise võtteid omavahel võrreldud. Seega täidab siinne bakalaureusetöö ühe tühimiku eesti lastekeelekorpuse arendamises. Eesti lastekeelekorpus suureneb pidevalt ning edasise uurimise huvides on oluline, et alamkorpused oleksid ühtlaselt koostatud. Siinse töö tulemusi ning tulemuste põhjal antud soovitusi võiks arvesse võtta edasisel korpuse koostamisel, see tagab korpuste koostamise ühtluse ja seeläbi ka andmete usaldusväärsuse.

## Kirjandus


**Argus**, Reili 2003. Lastekeelest ja selle uurimisest. − Oma Keel 1, 26−32. http://www.emakeeleselts.ee/omakeel/2003_1/OK_2003-1_03.pdf (24.05.2016).

**Argus**, Reili 2007. Eesti lastekeelekorpuse morfoloogilisest märgendamisest. – Toim. Pille Eslon. Tallinna Ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine. Tallinn: Tallinna Ülikooli Kirjastus, 65−86. http://evkk.tlu.ee/wwwdata/kogumik2007/korpusekogumik_argus.pdf (24.05.2016).

**Argus**, Reili 2008. Eesti lastekeelekorpuse morfoloogiliste vigade märgendamisest ja liigitamisest. – Toim. Pille Eslon. Õppijakeele analüüs: võimalused, probleemid, vajadused. Tallinn: Tallinna Ülikooli Kirjastus, 11−30. 4 http://evkk.tlu.ee/pdfs/kogumik2_2008.pdf (24.05.2016).

**Clark**, Eve V. 2009. First Language Acquisition. Second Edition. Cambridge: Cambridge University Press.

**Hennoste**, Tiit 2000a. Sissejuhatus suulisesse eesti keelde. − Akadeemia 10, 2223−2254.

**Hennoste**, Tiit 2000b. Sissejuhatus suulisesse eesti keelde. − Oma Keel 1, 48−57. http://www.emakeeleselts.ee/omakeel/2000_1/OK_2000-1_09.pdf (24.05.2016).

**Hennoste**, Tiit 2002. Keelekasutuse uurimine. − Emakeele Seltsi aastaraamat 48, 217−262. http://www.emakeeleselts.ee/esa/ESA_48_pdf/keelekasutus.pdf (24.05.2016).

**Hennoste jt** 2009 = Hennoste, Tiit, Olga Gerassimenko, Riina Kasterpalu, Mare Koit, Andriela Rääbis, Krista Strandson 2009. Suulise eesti keele korpus ja inimese suhtlus arvutiga. − Eesti Rakenduslingvistika Ühingu aastaraamat 5, 111−130. http://www.rakenduslingvistika.ee/ul/files/Hennoste-et-al_ERYa5_pp111-130.pdf (24.05.2016).

**Maarits**, Ethel 2011. Kõnesalvestuste transkribeerimine laste kõne korpuse näitel. Transcription of Speech on the Base of Children's Speech Corpus. Bakalaureusetöö. Tallinna Ülikool, informaatika instituut. www.cs.tlu.ee/teemad/get_file.php?id=91 (24.05.2016)

**MacWhinney**, Brian 1995. The CHILDES System; http://webcache.googleusercontent.com/search?q=cache:6YHx-m-7NjwJ:citeseerx.ist.psu.edu/viewdoc/download%3Fdoi%3D10.1.1.11.8342%26rep%3Drep1%26type%3Dpdf+&cd=1&hl=et&ct=clnk&gl=ee (16.05.2016).

**MacWhinney**, Brian 2000. The CHILDES Project. Tools for Analzing Talk. Third Edition. London: Lawrence Erlbaum Associates, Publishers.

**MacWhinney**, Brian 2001a. From CHILDES to TalkBank; http://repository.cmu.edu/cgi/viewcontent.cgi?article=1180&context=psychology (24.05.2016).

**MacWhinney**, Brian 2001b. The CHILDES System. − American Journal of Speech-Language Pathology; https://www.researchgate.net/profile/Brian_Macwhinney/publication/2551600_The_CHILDES_system/links/54e650070cf277664ff500d4.pdf (24.05.2016).

**Muischnek jt** 2003 = Muischnek, Kadri, Heili Orav, Heiki-Jaan Kaalep, Haldur Õim 2003. Eesti keele tehnoloogilised ressursid ja vahendid: arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Haridus- ja Teadusministeerium, Eesti keele nõukogu. Tallinn: Pakett. www.hm.ee/index.php?popup=download&id=3993 (24.05.2016).

**Parm**, Sirli 2013. Eesti keele ajasõnade omandamine. Dissertationes Linguisticae Universitatis Tartuensis 17, Tartu: Tartu Ülikooli Kirjastus. https://dspace.ut.ee/bitstream/handle/10062/29507/parm_sirli.pdf?sequence=1&isAllowed=y (24.05.2016).

**Tomasello**, Michael 2003. Constructing a Language. A Usage-Based Theory of Language Acquisition. London: Harvard University Press.

## Summary. Notation of Estonian Child Language in CHILDES-system

The research area of this study is language acquisition. This study focuses on the international child language database CHILDES and it's Estonian subcorpora.

Mostly, studies of language acquisition are based on experimental data or data collected from corpora. In addition to linguists, there are many other scientists that investigate child language development, for example psychologists, educationists, neurophysiologists and special education teachers (Argus 2003: 27).

So far, the Estonian subcorpora is made by child language researchers, who has recorded child speech and literated the data for their own research. Literation in Estonian subcorpora is based on researcher's personal interests (Argus 2007: 66).

The purpose of this study is to examine the differences between corpora and what causes these differences in transcription. I chose three recordings from each corpus. In case there were recorded more than one child in one corpus, I chose one recording for each child.

For this analysis I used the collected data from CHILDES' Estonian subcorpora and questionnaire I sent to compilers of these subcorpora. I compared transcriptions of comments, explanations, numbers, feedback words and foreign words.

The results show that there are some differences between corpora. Estonian subcorpora are increasing continuously and for further research it is important to follow identical transcription in every subcorpus.

# Lisad

## Lisa 1. Küsimustik korpuse koostajatele.

Küsimustik

1. Kellele ja miks on litereerimispõhimõtted vajalikud?
2. Millised on olnud Teie jaoks suurimad probleemid korpuse koostamisel?
3. Kuidas olete transkribeerinud, kui kaks iseseisvat sõna on kokku hääldatud?
4. Mida olete silmas pidanud järgmiste sõnade transkribeerimisel?
   - Üneemid *(näide*: *aa, ee, õõ)*
   - Tagasisideüneemid *(näide: mhmh, mqm, mh)*
   - Tagasisidesõnad *(näide: jaa)*
5. Kuidas olete transkribeerinud (märgistanud) ebaselgused kõnevoorus, sõnad vm?
6. Mille järgi otsustate, et algab uus kõnevoor?

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina, Paula Helena Kask
(sünnikuupäev: 02.06.1994)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Eesti lastekeele andmete esitus andmepangas CHILDES“, mille juhendaja on PhD Sirli Zupping,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, 24.05.2016