TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Kristiina Vaik

Morfoloogiliselt märgendatud eesti lapsekeele korpus

Magistritöö

Juhendajad: Heiki-Jaan Kaalep ja
Virve-Anneli Vihman

TARTU 2016

Sisukord

Sissejuhatus

See, kuidas lapsed keelt omandavad, on kognitiivteadustes olnud üheks keskseks uurimisvaldkonnaks. Keel on väga kompleksne süsteem, kuid juba varases eas on lapsed sellegipoolest võimelised lühikese aja jooksul omandama keele fonoloogilisi ja grammatilisi struktuure ning semantilisi ja pragmaatilisi suhteid. Aga see, kuidas lapsed seda teevad, tekitab siiani teadlaste seas palju vastakaid arvamusi.

Esimesed lapsekeele uurimise tööd põhinesid päevikumärkmetel, kus lapsevanemad dokumenteerisid oma lapse grammatika ja leksikoni arengut. 1940. ja 1950. aastatel hakati lapsekeele andmeid koguma süstemaatilisemalt, st hakati jälgima suure hulga laste keelelist arengut. 1960. aastatel ilmusid esimesed longituuduurimused, mis jälgisid lapse keelelist arengut teatud vaatlusperioodi vältel. Tehnoloogia areng mõjutas naturalistlike keeleandmete kogumise viisi: nii lapsevanemad kui ka keeleuurijad hakkasid andmeid koguma lapse spontaanse kõne lindistamise ja transkribeerimisega, mis sillutas teed suurte andmekogude ja uute uurimisküsimuste tekkimisele. Andmekogud võimaldasid uurijal süstemaatiliselt dokumenteerida ja analüüsida komplekssemaid keelelisi nähtusi nii lapse kui ka lapsele suunatud keeles, kuid need andmed olid kättesaadavad vaid väiksele hulgale uurijatele. Arvutitehnoloogia areng oli lapsekeele uurimises suureks edusammuks, sest nii suurenes andmekogude kättesaadavus. (Behrens, 2008) Kuid korpuste kättesaadavusega kerkis esile uus probleem: keeleomandmise uurimises puudus tol ajal asjakohane transkribeerimissüsteem, vt (Ochs, 1979).

1984. aastal lõid Brian MacWhinney ja Catherine Snow arvutipõhise andmebaasi CHILDES (*Child Language Exchange System*), mis nõudis andmete digitaliseerimiseks standartset süsteemi. Paljusid Ochsi ettepanekuid implementeeriti CHAT käsiraamatus (*Codes of the Human Analysis of Transcripts*), mis annab ülevaate CHILDES-i formaadi koostamise printsiipidest. CHILDES võimaldas keeleuurijatel oma keeleandmeid jagada, standartsel viisil transkribeerida, töödelda ning teiste keeltega võrrelda (MacWhinney & Snow, 1985). Lisaks CHAT käsiraamatule on keeleuurijatel võimalus kasutada CLAN tarkvara (*Computerized Language Analysis*), mis abistab keeleuurijat korpuse transkribeerimisel, kodeerimisel ja analüüsimisel.

Käesoleva magistritöö eesmärk on luua eesti morfoloogiliselt märgendatud lapsekeele korpus. CLAN tarkvara ei võimalda eesti keele jaoks teha muutevormide automaatset statistikat, sest süsteemis pole eesti keelele rakendatavat morfoloogilist analüsaatorit, mistõttu teevad lapsekeele uurijad hetkel distributiivset analüüsi käsitsi. Morfoloogiliselt märgendatud korpuse abil tekiksid uued võimalused uurimaks nii lapse kui ka lapsele suunatud kõne. Eesti lapsekeele korpuse struktuuri

esitamiseks ja morfoloogilise tasandi lisamiseks kasutasin oma loodud töövahendeid, morfoloogilist analüüsi teostasin eesti keele morfoloogilise analüsaatori *etana* abil. Loodud korpus pole ideaalselt analüüsitud, sest tegemist on esimese katsetusega. Töö käigus arutlen lapsekeelekorpuse transkribeerimisega seotud probleemidest ja muudest kitsaskohtadest, mis omakorda mõjutavad morfoloogilise analüüsi kvaliteeti.

Töö esimeses peatükis annan põgusa ülevaate lapse- ja hoidjakeele morfoloogiast. Teises peatükis räägin lähemalt korpuses olemusest, selle liigitusvõimalustest ja arengutendentsidest ning eesti keele korpustest. Lisaks annan lühiülevaate korpuse märgendamisest, morfoloogilisest analüüsist ja XML-ist. Kolmandas peatükis tutvustan CHILDES-i andmebaasi ja trankriptsioonisüsteemi ning tutvustan eesti keele korpuse struktuuri. Peatükis annan ülevaate ka alamkorpuste standardiseerimise probleemidest. Töö neljanda peatüki esimeses alajaotuses kirjeldan morfoloogiliselt märgendatud lapsekeele korpuse tööprotsessi: kirjeldan töö jooksul tekkinud probleeme, programmi töövoogu ja morfoloogilise info lisamist, ja teises alajoatuses hindan morfoloogilise märgenduse adekvaatsust. Viimases peatükis annan ülevaate sellest, kuidas korpusega edasi toimida ehk kuidas korpuse märgendamist ja standardiseerimist paremaks muuta ning kuidas oleks võimalik täiustada morfoloogilise analüüsaatori töö tulemust.

1. Lapse- ja hoidjakeele morfoloogia

Lapsekeel on keel, mida produtseerib laps ise ning mis aja jooksul muutub ja täiustub, sarnanedes lõpuks täiskasvanu kõnele. Lapsekeele all mõeldakse üldjuhul väikelapse kõnet, kuid konkreetseid vanuselisi piiranguid pole seatud. Keelekasutuse areng on individuaalne ja pidev, misõttu ei ole võimalik defineerida kindlat ajapiiri, mil enam pole tegemist areneva keelega. Hoidjakeel on lapsele suunatud keel ehk sisendkeel või -kõne. Lapse- ja hoidjakeelt võib vaadelda kui suulise keele allkeeli. Ehkki mõlemale keelele on iseloomulik mitteformaalne kõne ja emotsionaalselt lähedane kõnelemise situatsioon, tuleks neid kahte eristada, sest neil mõlemal on oma kindlad tunnused. (Kõrgesaar & Kapanen, 2015, 178–181)

Keeleomandamise uurimisel on pälvib enim tähelepanu see, kuidas omandatakse grammatikat. Vastuseid otsitakse küsimustele, kuidas ja millal grammatikat omandatakse, missugused tegurid mõjutavad morfoloogia omandamist ja kuidas mõjutab morfoloogia omandamine teiste keeletasandite omandamist. (Argus, 2008a, 10) Keeleomandamiskäsitlused võib jagada formaalseteks ehk generatiivsest grammatikast lähtuvateks ja kasutuspõhisteks lähenemisteks. Generatiivne lähenemine väidab, et laps analüüsib sisendkeelt lähtuvalt sünnipärastest kategooriatest, vt (Wexler & Culicover, 1980). Kasutuspõhise lähenemise järgi konstrueerivad lapsed grammatika sellest, mida nad kuulevad. Näiteks, lapse keelekasutusse tekkinud kindlad verbid on seotud nende verbide esinemissagedusega sisendkeeles. Lapsel tekib sõnavormidest fonoloogiliselt ja semantiliselt jagatud võrgustik, milles tekivad teatud paradigmad (nt nimisõnad, mis on sama käändelõpuga markeeritud, ja ühe nimisõna eri muutevormid). Produktiivsus ongi uue üksuse ja olemasoleva võrgustiku suhte tulemus. Produktiivsus sõltub tüübisagedusest ehk kui palju sõnu on sama mustri abil tuletatud ja piirangutest ehk mustri jagatud tunnuste mõjust uuele üksusele. (Lieven, 2010, 2546–2547, 2550; vt ka Krajewski et al., 2012)

Morfoloogiline paradigma piirdub alguses väikese arvu sarnaste vormidega, kuid järk-järgult muutub skeem abstraktsemaks ja produktiivsemaks (ehk mida vähem on jagatud tunnuseid, seda rohkem on see uutele vormidele rakendatav) (Lieven, 2010, 2549). Inglise keele verbi *go* eri vormide (ja ka tähenduse) omandamine võtab aega ning on seotud nende sagedusega sisendkeeles. Peale sageduse on ka muid võimalusi, nt mõned vormid on esilduvamad, prototüüpilisemad või fonoloogiliselt (+ morfoloogiliselt, semantiliselt) lihtsamad kui teised vormid, nt *go*, *going* vs *went*. (Theakston et al., 2002; vt ka Aguado-Orea & Pine, 2015) Eesti keeles on *minema* verbi omandamist uurinud Kaisa Seene (vt Seene, 2015).

Katsed laste ja täiskasvanutega kinnitavad, et kuigi muutemorfoloogia omandamist alustatakse varakult, siis muutemorfoloogia produktiivsus sõltub lisaks fonoloogi-

listele ja semantilistele faktoritele ka tüübi- ja sõnesagedusest. Kuid morfoloogia omandamise uurimisel ei piisa sellest, kui me lihtsalt loendame sisendkeeles esinevaid vorme ja vaatame, kuidas need peegelduvad lapse keelekasutuses. Lieven toob näite lapse poolt produtseeritud lausungitest, kus markeeritud verbi asemel kasutatakse selle markeerimata vormi. Hispaania keelt kõnelevad lapsed teevad selles osas vähem vigu kui hollandi või saksa keelt kõnelevad lapsed. Kasutuspõhise lähenemise järgi võiks mõelda, et kui sisendkeeles on markeerimata vormide suhteline sagedus suur, siis tehakse rohkem vigu markeerimise ärajätuga. Kuid tegelikult tuleb ilmsiks, et hispaania sisendkeeles on markeerimata verbivorme pea sama palju kui saksa või hollandi keeles. Niisiis tuleb kohe paika panna, mida täpselt sisendkeeles mõõta ja kuidas see suhestub lapse keelekasutusega. (Lieven, 2010, 2552–2554)

Eesti keele morfoloogia on väga mitmekesine, sest lisaks reeglipärastele mallidele tuleb omandada ka ebareeglipärased mallid, tuleb teada, kuidas toimuvad tüvesisesed muutused (astme- ja lõpuvaheldus), ning milliseid morfoloogilisi formatiive (tunnused ja lõpud) tüve külge lisada. Eesti keele vormimoodustust on palju uurinud Reili Argus (Argus, 2008a), kes samuti lähtub kasutuspõhisest lähenemisest. Hoidjakeelele on iseloomulik reduplikatsioon, mis hõlbustab kõnejada segmenteerimist ja silbipiiri ära tundmist (nt *ta-da*, *ai-ai* jne). Reduplikatsioon abistab last ka sõnade ära tundmisel ja aitab mõelda sõnadest kui mitmest osast koosnevast tervikust, mis omakorda hõlbustab hilisemas etapis tajuda sõnavormis esinevat muutumatut (tüvi) ja muutuvat (tunnused ja lõpud) osa. (Argus, 2008a, 19–20)

Reili Argus (Argus, 2008a, 23) kirjutab, et võiks justkui eeldada, et morfoloogilised formatiivid ja õiges astmes välte valimine raskendab morfoloogilise süsteemi omandamist, aga selgub, et eesti keeles omandatakse produktiivsed vältevaheldusmallid (nõrgeneva tüvega ühesilbilised substantiivid) juba perioodil, mil ei ole tunnused ega lõpud veel omandatatud. Vältevaheldus omandatakse varakult, kuna vältevaheldusliku sõnad on sisendkeeles sagedased ning väldete opositsioonidel on grammatiliste tähenduste eristamisel tähtis roll. Näiteks nõrga- või tugevaastmelised sõnad eristavad lapse jaoks grammatilisi tähendusi- valdaja ja objekt või objekt ja asukoht. Lõpuvaheldus pole nii süsteemne kui astmevaheldus ja selle omandamine on raske, kuna tüvevahelduslike vormide puhul tuleb lõpufoneemide järjestust vahetada. Probleeme valmistavad näiteks fonoloogiliselt keerulised sõnad, kus laps väldib kolmest konsonandist koosnevat ühendit (**numbert*, **numberit* '*numbrit*'; **kahvelga*, **kahveliga* '*kahvliga*') ja *s*-lõpulised sõnad (**kärbese* '*kärbse*'; **võõraseid* '*võõraid*'). Lihtsaim viis sõnade moodustamiseks on lõpuhäälikute lisamine, kuid vahel üldistatakse muutesufiksit ka sõnadele, kus see ei ole normikohane, nt *kauss*: **kausi-t* (partitiivi läbipaistva lõpu -*t* üldistamine), *tühi*: **tühja-sse* (illatiivi läbipaistva lõpu -*sse* üldistamine). (Argus, 2008a, 23–24, 26–27; Argus, 2008b, 20–21)

Oma bakalaureusetöös uurisin laste üldistamisvõimet (Vaik, 2014). Väljamõeldud stiimulitega moodustuskatsest selgus, et lapsed oskavad juba varases eas uutele sõnadele reegleid üldistada, kuid seda ei tehta veatult. Selgus, et lapsed on omandanud sufiksite reeglipärase lisamise, kuid probleeme valmistasid tüveteisendused. Tüveteisenduslikud vead olid seotud foneemi/silbi asendamise, lisamise või ärajätmisega, nt sõnade *lumber* ja *muhkur* (tüüpsõna *number*) puhul moodustati **lumberi*, **muhkuri*, **muhkurise*; sõnade *pülaline* ja *bobune* käänamisel moodustati **bobut*, **bobunet*, **pülalit*. Tüveteisenduslikest vigadest tehti enim laadivahelduslikke vigu, nt sõnade *päsi* ja *mäsi* (tüüpsõna *käsi*) puhul üldistati partitiivi läbipaistvat lõppu **päsit*, **mäsit*; sõnade *tammas* ja *mammas* (tüüpsõna *hammas*) puhul moodustati **tamma*, **mamma*, **tammase*, **mammase*. (Vaik, 2014) Need eksimused pole juhuslikud, vaid peegeldavad lapse arusaamu vormimoodustussüsteemi mustritest ja mallidest.

Eesti keeles on noomenitel lõpuvaheldusmalle rohkem kui verbidel, mistõttu ei valmista Arguse sõnul verbide lõpuvaheldusmallid lastele ka probleeme, näiteks sellised tüvevaheldused nagu *sööme*: *süüa*, *lööb*: *lüüa*, *ei pea*: *pidime* on omandatud veatult (Argus, 2008a, 24). Verbide omandamine sõltub sellest, kuidas laps neid sisendkeeles kuuleb. Vihman ja Vija uurisid verbi vormimoodustust kahe lapse keelekasutuses (vt Vihman & Vija, 2006), kuid vaatluse alla võeti lisaks ka ühe lapse hoidja keelekasutus. Mõlemad lapsed kasutasid juba varases eas markeerimata verbivorme, mis olid ka sisendkeeles ühed kõige sagedasemad (imperatiivi ainsuse 2. pööre ja oleviku vormi eitus). Lisaks oli mõlema lapse puhul oli märgata, et ebaregulaarseid verbivorme omandatakse kiiremini kui reeglipäraseid, sest nende esinemissagedus on sisendkeeles suur. Verbide moodustamisel tegid lapsed kolme sorti vigu: markeerimata verbi kasutamine (*mul on vaja *pese kätt*, *valge kiisu *tudu*); fonoloogilised vead (*ma *süüan*, **leidsi ülesse*, **oleb*, **loes*, **sajas*) ja morfoloogilised vead (*ei taha putukas *läheme sisse*, *sina *on paha poiss*). (Vihman & Vija, 2006)

Keeleomandamise varases etapis on hoidjakeelele iseloomulik eripärane intonatsioon, lühemate lausete (ka verbita lausete) kasutamine, reduplikatiivsete ja onomatopoeetiliste sõnade rohkus, uue info kordamine, keerukate muutevormide vältimine, küsilaused, mitmuse esimese isiku ja deminutiivtuletiste kasutamine. (Kõrgesaar & Kapanen, 2015, 178–182; Orusalu, 2008, 26–29) Kuid lapse vanuse kasvades muutuvad lausungid lapsele suunatud kõnes pikemaks, onomatopoeetilised sõnad ja deminutiivtuletisted hakkavad kaduma, küsilauseid ja korduseid jääb vähemaks (Kõrgesaar, 2009, 37; vt ka Vihman, 2015). Hoidjakeel (kui ka muu ümbritsev keel) on keeleomandamise seisukohast oluline, sest sisendkeel hõlbustab lapsel sagedasemate keeleüksuste omandamist ja muudab lapse sisendkeele statistiliste ja süntaktiliste tunnuste suhtes vastuvõtlikumaks. Paraku pole kõik hoidjakeelele ise-

loomulikud tunnused (nt fonoloogiline lihtsustamine, reduplikatiivsete sõnade ja deminutiivtuletiste kasutamine) lapse jaoks ilmtingimata kõige kasulikumad, sest sisendi keelelise vaesuse tõttu peab laps puuduolevad üksused ise rekonstrueerima. (Soderstrom, 2007; vt ka Newport et al., 1977)

Morfoloogiasüsteemi omandamine toimub sõnaliigiti erineva kiirusega. Arvatakse, et noomeni morfoloogia omandatakse kiiremini kui verbi morfoloogia. Seda on põhjendatud sellega, et noomenite puhul tuleb omandada vähem morfoloogilisi kategooriaid. Lisaks arvatakse, et verbe omandatakse teisiti, sest noomenite referentsiaalsust on kergem hoomata ning verbid on semantiliselt keerukamad ja on rohkem seotud keele süntaktilise struktuuriga. (Gentner, 1982) Ometigi uurimused näitavad, et keeltes, kus hoidjakeeles esineb rohkem verbe, on ka lapse varases kõnes sõnaliigiti kõige sagedasem verb (vt korea keel (Choi & Gopnik, 1995); mandariini keel (Tardif, 1996)). Nimisõnad on ühtlasema jaotusega kui verbid, st nimisõnu on küll palju, kuid ükski nimisõna ei tõuse sageduse poolest esile, ja erinevaid verbe on küll vähe, kuid mõned neist on esilduvamad (nt *tegema* ja *olema*). Seega peab laps uue verbiga kokku puutumisel tegema rohkem üldistusi kui nimisõnadega. (Argus & Kõrgesaar, 2014, 38–39)

Sõnaliikide jaotumist eesti keeles on uurinud Kadri Vider (Vider, 1995) ning Argus ja Kõrgesaar (Argus & Kõrgesaar, 2014). Kadri Vider uuris, kuidas sõnaliikide esinemissagedust lapsekeeles ja tema andmed pärinevad lindistustest lastega vanuses 1;11–3;11. Ta koostas sõnaliikide sagedussõnastiku, kus leksikaalsete üksuste tasandil leidus kõige enam substantiive, sellele järgnesid verbid, adverbid, adjektiivid, interjektsioonid, pronoomenid, kaassõnad, numeraalid ja konjunktsioonid, kuid sõnavormide tasandil oli esinemissageduse poolest kõige enam verbe, järgnesid substantiivid, adverbid ja pronoomenid, konjunktsioonid, adjektiivid, interjektsioonid, kaassõnad ja numeraalid. (Vider, 1995)

Argus ja Kõrgesaar uurisid sõnaliikide esinemissageduste jaotumist nii lapse- kui ka hoidjakeeles, kuid lisaks traditsioonilisele sõnaliigi jaotumisele võeti eraldi arvesse ka onomatopoeetilised sõnad. Selle uurimuse põhjal võib väita, et eesti laste varajane kõne on nimisõnakeskne, kuid see ei tulene nimisõnade suurest sagedusest sisendkeeles, kuna vaatlusperioodi alguses oli nimisõnu ja verbe pea sama palju ja vaatlusperioodi lõpus oli verbide osakaal nimisõnadest suurem. Onomatopoeetiliste sõnade osakaal lapse- ja hoidjakeeles on erinev: lapsekeeles on neid vaatlusperioodi alguses küllaltki palju, kuid enne 2-aastaseks saamist hakkab see järk-järgult vähenema. Hoidjakeeles oli võrreldes lapsega vähe onomatopoeetilisi sõnu ja enne lapse 2-aastaseks saamist muutub nende osakaal pea olematuks. Hoidjakeelele oli iseloomulik ka väike adjektiivide, kaassõnade, konjunktsioonide ja numeraalide osakaal, mis omakorda peegeldus ka lapse keelekasutuses. (Argus & Kõrgesaar, 2014)

2. Korpused

Selles peatükis tehakse lühike kokkuvõte korpustest. Selgitatakse lähemalt mida mõeldakse korpuse ja muude oluliste mõistete all. Kirjeldatakse lühidalt ajaloolist tausta. Peatüki teine osa annab ülevaate eesti keele korpustest. Viimasena keskendutakse korpuse märgendamisele ja sellega seonduvatele olulistele mõistetele.

2.1. Mis on korpus?

Enne arvutite kasutuselevõttu mõeldi keeleteaduses keelekorpuse all keelekogumikku, mida sai keeleuurija (vastandina enda intuitsioonile) kasutada uurimustöö algmaterjalina. Tänapäeval mõistetakse keelekorpuse all elektroonilisel kujul olevat tekstikogu, kuhu lisatakse tekste eesmärgiga, et need annaksid tõepärase pildi keelest ja iseloomustaksid keele hetkeseisu või muutumist. (Muischnek et al., 2003, 9)

Sellist korpust, kus tekstid esindavad teatud ajavahemiku keelekasutust, nimetatakse representatiivseks ehk suletud korpuseks. Suletud korpuses ei saa tekste ära võtta ja juurde lisada. Suletud korpus ei pruugi teatud aja möödudes olla enam representatiivne, kuna keel ja selle sõnavara muutub. Avatud ehk monitorkorpuste puhul ei valita tekste rangete kriteeriumite alusel, talletada võib tekste, mis võivad vastata koguja vajadustele või mida on olnud võimalik (lihtsalt) koguda. Erinevalt suletud korpusest saab avatud korpusesse tekste alati juurde lisada. Lisaks avatud–suletud liigitusele, võib korpuseid liigitada ka mitmete teiste tunnuste alusel, nt kirjalik vs suuline (lisandunud on ka uue meedia ehk internetikeel); ükskeelne-kakskeelne-mitmekeelne; katkendikorpus vs tekstikorpus; diakrooniline vs sünkrooniline; allkeel vs üldkeel ja puhas tekst vs märgendatud tekst. (Muischnek et al., 2003, 9–11; McEnery & Hardie, 2011, 9–13)

Korpused jagunevad kolme põlvkonda:

1. põlvkond (ca 1960ndate lõpp–1980ndate lõpp): suletud, representatiivne, väike, valdavalt 80ndatel tehtud, palju käsitööd panustatud. Nt Brown, LOB, Frown, kirjaliku eesti keele 80ndate aastate korpus;

2. põlvkond (valdavalt 1990ndate teises pooles ja 2000ndatel): avatud, suured tekstihulgad, elektrooniliste publikatsioonide teisendamine ühtsele korpuse kujule. Nt eesti keele koondkorpus;

3. põlvkond: väga suur, automaatselt veebist korjatud ja ühtsele korpuse kujule

teisendatud, nt etTenTen. Miinused: palju sodi, pole täpselt teada, mida korpus sisaldab. (Muischnek, 2015, 37–38)

Esimesed elektroonilised tekstikorpused olid Browni ja Lancaster-Oslo/Bergeni (LOB) korpused. Nendesse korpusestesse lisatud tekstid olid väga läbimõeldud ja koosnesid vaid ühest miljonist sõnast, mis pole tänapäeva korpuste mahtudega võrreldav. Põhjuseks oli muidugi tehnikaareng. Browni ja LOB-i korpuste loomise ajal polnud arvutite jõudlus ja mälu nii suur, et suudaks talletada ja töödelda rohkemat kui üht miljonit sõna. Tänapäeval pole see enam probleemiks. Ligi 20 aastat olid nende korpuste koostamise põhimõtted olnud standardiks ka paljude teiste keelte korpuste loomisel, sealhulgas ka tänapäeva eesti kirjakeele baaskorpuse jaoks. Tänu tehnika arengule tekkisid võimalused suuremate tekstikorpuste loomiseks. Näiteks, 1991. aastal tehti Inglismaal algust kahe suure projektiga: *British National Corpus* (BNC) ja *Bank of English* (BoE). BNC on suletud korpus, mille maht on 100 miljonit sõna. BoE on avatud monitorkorpus. BoE on mõeldud eeskätt leksikograafidele kasutamiseks. (Muischnek et al., 2003, 9–11)

2.2. Eesti keele korpused

Tänapäeva kirjakeele korpus sai alguse 80ndate aastate *baaskorpusest*, mille standardiks on Browni ja LOB-i korpused. Eesti kirjakeele korpus on suletud ja representatiivne. Korpus koosneb ühest miljonist sõnast, tekstid pärinevad aastatest 1984–1987 ja on jaotatud kümnesse tekstiklassi. Baaskorpusega liituvad ka *niitkorpused* ehk *läbilõikekorpused* (1890–1990), mis on suletud ja osaliselt representatiivsed, kuigi neis on vähem tekstiklasse. Baas- ja läbilõikekorpustes on kokku umbes 4 miljonit sõna. (Muischnek et al., 2003, 14–15) *Koondkorpus* (sai alguse 1990ndatel) on teise põlvkonna avatud korpus, mis koosneb umbes 250 miljonist sõnast. Korpus sisaldab palju ajalehetekste ja kasutatakse terviktekste (mitte katkendeid). Koondkorpuse alamhulk on *Tasakaalus korpus*, mis sisaldab 5 miljonit sõna nii ilukirjandust, ajakirjandustekste ja teadustekste. Kolmanda põlvkonna korpused on automaatselt veebist korjatud, sisaldades foorumite, blogide ja kommentaariumide tekste. (Muischnek, 2015, 38) etTenTen koosneb 270 miljonist sõnast 686000 veebilehelt. (etTenTen, 2015)

Kõik eelnevad sisaldavad kirjalikku eesti keelt, kuid koostatakse ka mitmeid erikorpuseid:

1. paralleelkorpus
2. suulise kõne korpus;

3. spontaanse kõne foneetiline korpus
4. dialoogikorpus;
5. murrete korpus;
6. vana kirjakeele korpus;
7. foneetikakorpus jne.

(*Korpused ja keelekogud*, 2015; Muischnek et al., 2003, 17–22)

2.3. Korpuse märgendamine

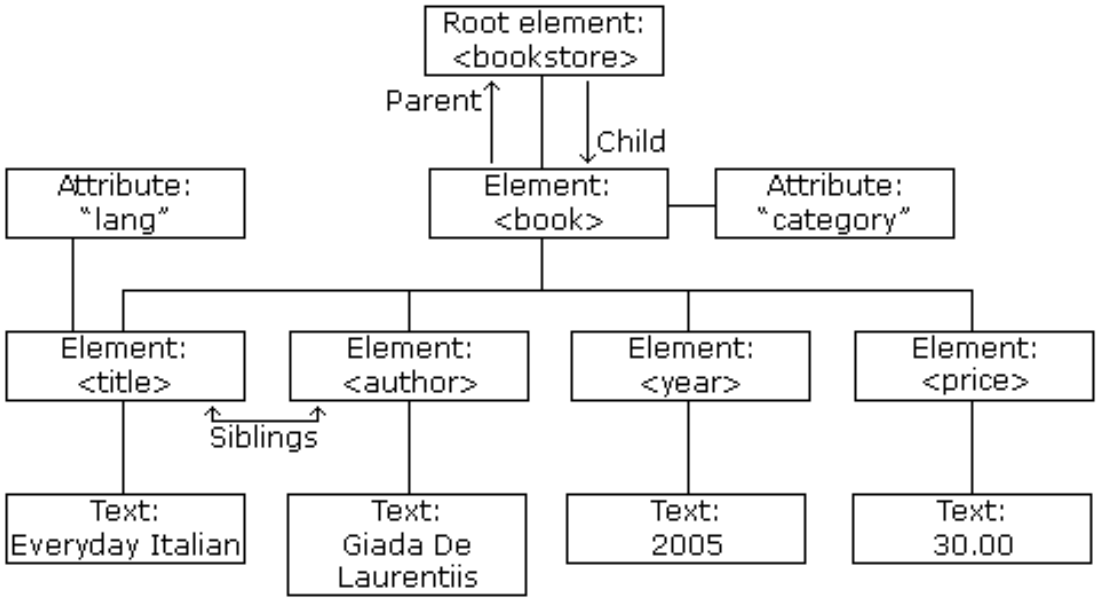
Korpustest on kasu siis, kui vajalik info on sealt lihtsasti kättesaadav. Selleks, et korpus ei jääks lihtsalt elektrooniliste tekstide arhiiviks, oleks tarvis korpusesse esmalt interpretatiivset infot lisada. (Muischnek et al., 2003, 12) Korpus hõlmab endas (tüüpiliselt) kolme tüüpi informatsiooni: metaandmed, teksti elementide märgendus (*corpus markup*) ja lingvistiline märgendus (*annotation*). (McEnery & Hardie, 2011, 30; vt ka Burnard, 2014)

Metaandmed sisaldavad informatsiooni teksti enda kohta – nt autor, aeg, keel, osalejad, vanus, sugu, kontekst jne. Lingvistilise märgenduse etapil on vajalik selgeks teha, mida (sisu) ja kuidas (vorm) märgendada. Märgendamist saab teha automaatselt, käsitsi või neid mõlemaid kombineerides (poolautomaatselt). Lingvistilist märgendamist alustatakse esmalt tehnilise märgendamisega: lausestaja abiga pannakse paika pealkirjad, autorid, lõigud, laused, tabelid ja väljajäetud materjal. Seejärel tuleb valida vastavalt korpuse eesmärgist märgendustase(med), nt morfoloogiline, süntaktiline, semantiline, pragmaatiline märgendus. Oluline on see, et korpus oleks korralikult ja standardselt märgendatud, sest nii saab seda korpust kasutada uuesti erinevate eesmärkide tarbeks. Näiteks morfoloogiline analüüs on kõikide teiste märgendustasemete (süntaktiline, semantiline jne) alus. (Muischnek et al., 2003, 12–14)

Teksti elementide märgendus kodeerib tekstisisest informatsiooni. Näiteks seda, millal kõneleja kõnevoor algab ja lõpeb. Märgendamisel on tähtis säilitatada teksti algandmete kohta võimalikult palju infot ja et see oleks inim- ja masinloetav. Korpuse teksti elementide märgendamise ühe viisina kasutatakse XML-i ehk *EXtensible Markup Language*. (McEnery & Hardie, 2011, 29–30) Järgnevalt tutvustangi töö seisukohalt olulisi teemasi: XML ja morfoloogiline analüüs.

2.3.1. Mis on XML?

XML (*EXtensible Markup Language*) on *World Wide Web* konsortsiumi poolt soovitatud markeerimiskeel, mille eesmärk on andmete talletamine ja jagamine erinevate infosüsteemide vahel. XML-dokumentides kujutatakse andmeid hierarhilise puustruktuurina. XML-i puu koosneb juurelemendist (*root*), millel on alamelemendid ehk järglased (*child elements*). Kõikidel elementidel võib olla järglaseid. Elementidevahelisi suhteid kirjeldavad sellised mõisted nagu ülem (*parent*), alluv (*child*) ja kolleeg (*sibling*). Ülemal on alluvad, alluval on ülem ja kolleegid on samal tasemel paiknevad alluvad. Kõikidel elementidel võib olla sisu (*text content*) ja atribuut ehk tunnus, mis täpsustab või kitsendab elementi. (*XML Tutorial*, 2016) Joonis 1 illustreerib raamatupoe elementide ehk raamatute hierarhilist struktuuri:



Joonis 1: Raamatupoe hierarhiline struktuur (*XML Tutorial*, 2016)

Joonise 1 kujutamine XML-kujul:

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
  <book category="cooking">
    <title lang="en">Everyday Italian</title>
    <author>Giada De Laurentiis</author>
    <year>2005</year>
    <price>30.00</price>
```

```
  </book>
  <book category="children">
    <title lang="en">Harry Potter</title>
    <author>J K. Rowling</author>
    <year>2005</year>
    <price>29.99</price>
  </book>
  <book category="web">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
  </book>
</bookstore>
```

(*XML Tutorial*, 2016)

XML-i võib vaadata kui reeglite kogumikku, milles talletatakse informatsiooni semantiliste märgendite abil. Märgendid (*tags*) on <> märkide vahel olevad muutujad ja igal märgendil peab olema lõpumärgend (nt *<bookstore>* ja *</bookstore>*). XML dokument koosneb kolmest osast: proloog, dokumendi element ja epiloog. Faili alustatakse proloogiga, mis defineerib XML-i versiooni ja kasutatava kodeeringu. Dokumendi element on juurelement, mida saab olla vaid üks. Joonise 1 juurelement on *<bookstore>*, mille alluvaks on elemendid *<book>*. Märgenditel võib olla atribuut kui ka sisu, kuid need pole ilmtingimata kohustuslikud. Selles XML-koodijupis on raamatutel defineeritud ka atribuut *category*, mille väärtus oleneb raamatu valdkonnast. Elemendi *<book>* alluvateks on *<title>*, *<author>*, *<year>* ja *<price>*, mis on omakorda teineteise kolleegid. Kõigil neil elementidel on sisu ja elemendil *<title>* on ka atribuut *lang*, mille väärtuseks on keel. Viimane rida (*</bookstore>*) ütleb, et see on juurelemendi lõpp ja ühtlasi ka dokumendi keha lõpp. See tähendab, et rohkem raamatuid selles raamatupoes ei eksisteeri. (*XML Tutorial*, 2016)

Märgendite abil pannakse paika andmete loogiline struktuur. XML-il pole eeldefineeritud märgendeid. Seega igal inimesel on võimalik defineerida oma vajadustele vastav struktuur ehk süntaks, mis paneb paika elemendi nimetused ja järjestuse. Oluline on, et kasutaja poolt defineeritud süntaks vastaks XML-i rangetele reeglitele:

1. eksisteerib juurelement;
2. elementidel peab olema lõpumärgend;

3. elementide pesitsemine ehk üksteise sees paiknemine (*nesting*) on rangelt määratletud;

4. atribuutide väärtused peavad olema jutumärkides. (*XML Tutorial*, 2016)

Kui kasutaja loob enda märgenduse, siis XML-protsessoril pole võimalik selle valiidsuses veenduda, sest pole midagi millegagi võrrelda. Selleks tuleb kasutajal XML-dokumendis defineerida kasutatav süntaks. XML-dokumentide valideerimiseks on kaks viisi: dokumenditüübi definitsioon (*document type definition* – DTD) ja XML-skeema (*XML schema*). Nende asukoht on vahetult peale XML-versiooni deklaratsiooni ja kindlasti enne dokumendi keha. Juhul kui XML-dokument on DTD või XML skeemaga vastavuses, siis on ka XML-dokument kehtiv. (*XML Tutorial*, 2016)

Sellisel markeerimiskeelel on küll palju eeliseid, kuid puudusteks peetakse seda, et see on verboosne (nt kohustuslik lõpumärgend), nõuab kõrget valideerimisstandardit ja elemendid peavad rangelt üksteise sees paiknema. Näiteks suulises kõnes on palju pealerääkimisi, kuid vastavalt XML-i rangetele reeglitele pole elementide ristumine võimalik. Oma XML-i süntaksi kirjutamine pole lihtne (kõigi eelnevalt nimetatud puuduste tõttu), seega eelistavad inimesed kasutada juba teada tuntud kodeerimisskeemasid (nt CHILDES, Brown jne). (Leech, 2005)

2.3.2. Morfoloogiline analüüs

Morfoloogilise analüüsi käigus lisatakse iga sõna jaoks infot selle lemma ehk algvormi, sõnaliigi ja morfoloogiliste kategooriate kohta: käändsõnal arv ja kääne, tegusõnal pööre, tegumood, aeg, kõneviis, kõneliik. Teksti morfoloogiliseks analüüsimiseks kasutatakse morfoloogilist analüsaatorit. Analüsaator on programm, mis saab sisendiks teksti ja mille väljundiks on morfoloogiliselt analüüsitud sõnad. (Kaalep & Vaino, 2000, 89)

Morfoloogiline analüüs koosneb kahest etapist – üksiksõnade analüüsimine ja ühestamine. Analüsaator annab igale sõnale selle interpretatsioonid ehk analüüsivariandid. Seejärel toimub morfoloogiline ühestamine. Morfoloogiliseks ühestamiseks nimetatakse protsessi, kus kõikvõimalikest interpretatsioonidest tuleb välja valida antud konteksti sobiv analüüs.(Kaalep & Vaino, 2000, 90)

Selles magistritöös kasutatakse lapsekeele korpuse morfoloogiliseks analüüsimiseks sõnastikupõhist morfoloogilist analüsaatorit. Sõnastikupõhisel morfoloogilisel analüüsimisel töödeldakse sõnavorme ja võrreldakse antud keele leksikoniga, juhul kui sõna leksikonis pole, siis kasutatakse mitmesuguseid heuristilisi reegleid. Sõnasti-

kus eristatakse sõnaliike järgnevalt:

- A = adjektiiv - algvõrre e positiiv;
- C = adjektiiv - keskvõrre e komparatiiv;
- D = määrsõna (e adverb);
- G = genitiivatribuut (käändumatu omadussõna);
- H = pärisnimi;
- I = hüüdsõna (e interjektsioon);
- J = sidesõna (e konjunktsioon);
- K = kaassõna (pre/postpositsioon);
- N = põhiarvsõna (e kardinaalnumeraal);
- O = järgarvsõna (e ordinaalnumeraal);
- P = asesõna (e pronoomen);
- S = nimisõna (e substantiiv);
- U = adjektiiv - ülivõrre e superlatiiv;
- V = tegusõna (e verb);
- X = verbi juurde kuuluv sõna, millel eraldi sõnaliigi tähistus puudub;
- Y = lühend;
- Z = lausemärk. (*Vabamorfi morfoloogia-leksikon*, 2016)

98% eesti keele sisendteksti sõnadest on analüüsitavad nii, et kasutatakse sõnastikku, morfeemide loendeid ja nende kombineerimise eeskirju. Sõnadel lõigatakse maha lõpud ja liited ning võrreldakse sõnastikus olevate lekseemidega. Kuni 3% teksti sõnadest pole sõnastikupõhiselt võimalik analüüsida, sest sõnastikus puudub selle kohta kirje. Oletamisega oletatakse sõna algvorm ja selle vorm puhtalt sõnavormi alusel. Paraku analüsaator ei paku alati õigeid analüüsivariante (vale analüüse kuni 0,1%) ja selle peamised vead seisnevad selles, et sisendtekst pole analüsaatori jaoks mõeldud (analüsaator on mõeldud kirjakeele analüüsimiseks) ja et pärisnimed sarnanevad vormilt üldnimisõnadega. (Kaalep & Vaino, 2000, 91–93)

3. CHILDES ja eesti keele alamkorpused

Selles peatükis tutvustatakse CHILDES-i korpust ja eesti lapsekeele korpuste struktuuri, mida siinkirjutaja kasutab magistritöö andmestikuna. Lisaks antakse lühiülevaade alamkorpuste standardiseerimise probleemidest.

3.1. CHILDES

CHILDES (*Child Language Data Exchange System*) on *Talkbank*i alamkorpus, mis loodi 1984. aastal Brian MacWhinney (Carnagie Melloni ülikool) ja Catherine Snow (Harvardi ülikool) poolt selleks, et koondada kokku erinevate keeleuurijate kogutud keelematerjali eesmärgiga, et need oleksid kõigile vabalt kättesaadavad ja võimaldaksid eri keelte uurijatel oma andmeid ja uurimistulemusi teiste keeltega võrrelda. CHILDES-ist on saanud mahukas, rahvusvaheline ja usaldusväärne andmebaas, mis sisaldab nii audio- ja videolindistusi kui ka standardsel viisil transkribeeritud tekste. (Gillis, 2014, 1)

CHILDES-i süsteemi juures peab silmas pidama seda, et see funktsioneerib *repositooriumina*. Repositooriumist võib mõelda kui laost või arhiivist, kuhu üles laetud materjali talletatakse digitaalselt. Repositooriumi jaoks on oluline, et korpused oleksid avalikult kättesaadavad ja standardsel viisil transkribeeritud ja et andmekogu oleks kooskõlas rahvusvaheliste standarditega. Seetõttu pakub CHILDES erinevaid tarkvaralisi töövahendeid, mida arendatakse ja kaasajastatakse kõigil platvormidel (*Windows*, *MacOS*, *Unix*). (Gillis, 2014, 1)

CHILDES-i andmebaas jaguneb nelja suurde kategooriasse:

1. esimese keele omandamine;
2. teise keele keele omandamine;
3. kakskeelsus ja
4. kliinilised probleemid. (Gillis, 2014, 1)

Andmebaasi tekstide/lindistuste transkribeerimiseks ja kodeerimiseks kasutatakse CHAT käsiraamatut (*Codes of the Human Analysis of Transcripts*, vt (MacWhinney, 2016)). CHAT käsiraamat on mõeldud selleks, et kõik lindistused/tekstid oleksid standardsel viisil transkribeeritud ja kodeeritud. Käsiraamatus on väga suur valik kodeeringuid, kuid transkribeerija ei ole kohustatud neid kõiki kasutama. Oluline oleks, et transkribeerimist ja kodeerimist tehakse vähemalt baastase-

mel. Lisaks CHAT käsiraamatule on keeleuurijatel võimalus kasutada ka analüüsimistarkvara ja redaktorit CLAN (*Computerized Language Analysis*), mis abistab keeleuurijat korpuse transkribeerimisel, kodeerimisel ja analüüsimisel. CLAN võimaldab analüüsida kollokatsioone, sõna- ja foneemisagedusi, arvutada vormide ja lausungite keskmisi pikkusi. CLAN tarkvaraga loodud failiformaati nimetatakse CHAT-failiks ja see salvestatakse laiendiga *.cha* (xxx.cha) (Gillis, 2014, 1–2, 6) Talkbankis on kasutusel ka Chatter tarkvara, mis teostab CHAT-failide ranget valideerimist ja ka konverteerimist valiidseteks XML-failideks (*Chatter tarkvara*, 2016).

Hetkel on andmebaasis esindatud 39 keelt ja 2013. aasta maikuu seisuga koosnes andmebaas 13 miljonist lausungist ja rohkem kui 50 miljonist sõnavormist. Kõige suurema mahuga on esimesse kategooriasse kuuluvad ehk esimese keele korpused (11 miljonit lausungit ja 43 miljonit sõnavormi). Kõige suurema esindatavusega on inglise, saksa ja prantsuse keel.(Gillis, 2014, 2–5)

Transkriptsioonid algavad päisega (ingl k. *header*), kus antakse informatsiooni lindistuse aja, koha, osalejate, kestuse, laste vanuse jms kohta. Põhiridadele paigutatakse kõnelejat tähistav kolmetäheline kood, millele järgneb kõneleja tegelik kõne. Tegelikule kõnele lisatakse juurde, kas transkribeerija- või uurijapoolsed kommentaarid või kodeeringud (neid nimetatakse *sõltridadeks*). Sõltridade arv oleneb keeleuurija eesmärkidest. Nagu suulise kõne puhulgi, pole ainuüksi verbaalse info järgi aru saada, millest hetkel jutt käib, seega tuleks transkribeerimisel kasutada vähemalt üht sõltrida, nt kommentaariridа. (Argus, 2007, 68; MacWhinney, 2016) Vt näide (1) ja (2).

(1):

*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.

(Kõrgesaar, gregory03.cha)

(2)

*FAT: köögis saab teritada , köögis on nuga .

*MOT: +< aga siin oli ka teritaja .

*FAT: jaa aga ma ei tea , kus see on .

*CHI: +< seda kätte .

*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.cha)

3.2. Alamkorpuste standardiseerimise probleemid

Reili Argus kirjeldab oma artiklis (Argus, 2007) mõningaid transkribeerimise ja CHILDES-i tarkvara kasutamisega seonduvaid probleeme. Esiteks, analüüsitarkvara CLAN on mõeldud inglise keelele, seega tuleb eesti keele analüüsimisel arvestada sellega, et eesti keel on võrreldes inglise keelega sünteetilisema süsteemiga keel. Seega, kui keeleuurija tahab CLAN-i kasutades teha mingisuguseid sagedusloendeid, siis ei saada adekvaatseid tulemusi. Näiteks lekseemi *kala* kolm sõnavormi *kala*, *kalaga*, *kalale* loetakse programmi poolt eri lekseemideks. Selline asjaolu põhjustab ka statistiliselt väärate arvude tekkimist. Homonüümide eristamist tuleb teha näiteks käsitsi.(Argus, 2007, 70)

Argus väidab, et kuna lindistuste transkribeerimisel kasutatakse kuuldeortograafiat, siis need transkriptsioonid ei anna tõetruud pilti sellest, milline on lapse tegelik keelekasutus. Kui juba suulise kõne automaatne analüüsimine on keeruline, siis on lapse suulise keele analüüsimine veel keerukam. Lindistuste puhul on tegemist spontaanse suulise kõnega, mis sisaldab elemente, mida pole tarvis analüüsida, nt häälitsused. Seega 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)

Teiseks, probleeme tekitab see, et lapse puhul on tegemist ju areneva keelekasutusega, milles esineb palju erilisi tunnuseid, nt sõnakordus. Näiteks, kui sellist lausungit transkribeeritakse nii **CHI: onu, onu, onu*, siis tähendab see seda, et lapse lausung koosneb kolmest sõnavormist, aga kui näiteks transkribeerida seda lausungit

viisil **CHI: onu [/] onu [/] onu [/]*, siis koosneb see lausung ühest sõnavormist, kuna CLAN tarkvara kohtleb seda kui korduvat üksust. Lisaks sõnakordusele on probleemiks ka onomatopoeetilised sõnad, mida esineb lapsekeeles väga palju ja seetõttu tuleb transkribeerimisel läbi mõelda, kuidas selliseid juhtumeid lahendada. CHAT käsiraamat soovitab onomatopoeetiliste sõnade lõppu lisada sümbol @ (Argus, 2007, 72–73), aga reaalsuses kasutatakse seda ikka väga vähe ja see omakorda põhjustab seda, et transkriptsioonid ei järgi ühtset märgendamisstiili.

Võrreldes täiskasvanutega esineb lapsekeeles rohkem vigaseid vorme. Transkribeerimisel (nt vigade korral) on oluline, et transkribeerija peab nägema ja teadma seda, mida tegelikult öelda taheti, ja vastavad kodeeringud ka transkriptsiooni lisama nii, et vead oleksid juba esimesel tasandil liigitatud. (Argus, 2007, 74) Kahjuks praegused alamkorpused pole veakodeerimise osas järjepidevad, kord on viga kodeeritud ühtmoodi, kord teistmoodi ja vahel üldse mitte. Näites (2) on viga põhireal kodeeritud kooloniga (*:*), mille järele lisatakse korrektne sõnavorm. Näites (3) on veakodeerimine hoopis teine: põhireal järgneb vigasele sõnale [*] ja sõltreale on lisatud vearida (*%err*), kus toimub vea lahtikodeerimine ja sümboli = järele lisatakse korrektne vorm. Näites (4) on viga üldse kodeerimata jäetud.

(2)

*FAT: kriit pane tahvli peale .

*CHI: kit [: kriit] .

*CHI: kit [: kriit] (.) vahvlile [= tahvlile] pääle [: peale] .

(Vija; 20007.cha)

(3)

*CHI: issi , loe seda .

*CHI: issi , nüüd see [*] ei pane kinni !

%err: see=seda $MOR

(Vija; 20007.cha)

(4)

*FAT: viskad minema või?

*FAT: kus sa viskad selle?

*CHI: kinn.

*FAT: sinna viskad jah.

(Kõrgesaar; arabella01f.cha)

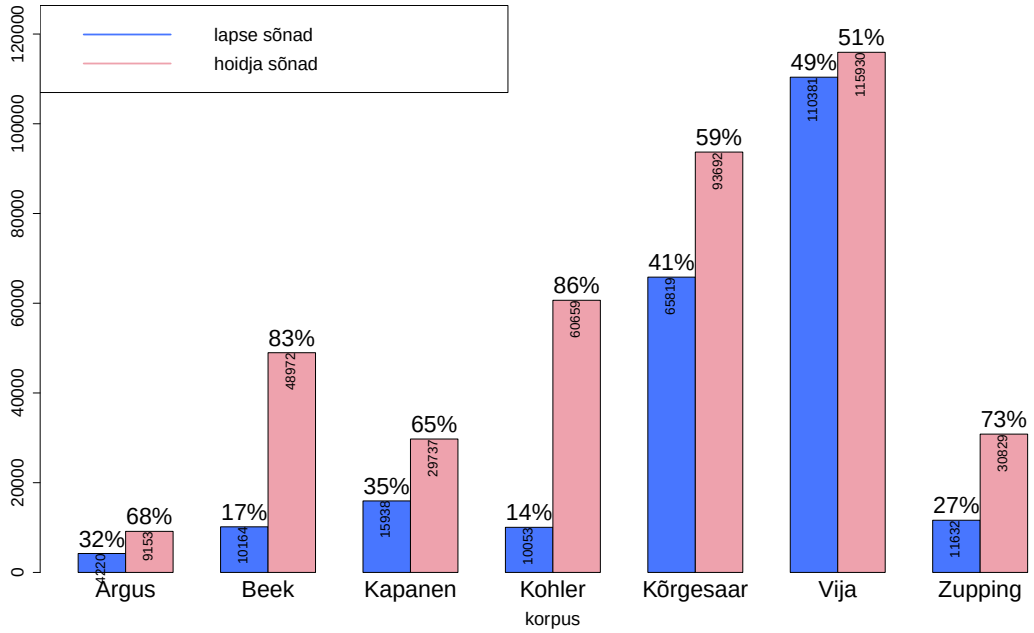
Morfoloogiliselt märgendatud korpuse loomine on väga vajalik, sest CHILDES-i analüüsitarkvara ei võimalda eesti keele morfoloogilist analüüsimist ja selle käsitsi tegemine oleks väga ajamahukas töö. Seega, praegusel hetkel on lapsekeele uurijatel automaatse statistika tegemine raskendatud ja paraku tehakse distributsioonianalüüse käsitsi (Argus, 2007, 78).

3.3. Eesti keele alamkorpuste struktuur

Selles töös kasutatakse kolm keskset mõistet – *sõna*, *sõnavorm* ja *sõnavara*. Sõna all mõeldakse tekstisõna ehk tühikute vahele jäävat tähtede järjendit, nt lauses *võtan teise pliiatsi* koosneb kolmest sõnast. Sõnavorm on unikaalne tekstisõna, nt lauses *pliiats ja pliiatsiga* on kaks sõnavormi (need on ühe ja sama lekseemi *pliiats* erinevad grammatilised vormid, siis selles töös käsitletakse neid kui kaht eri sõna); lauses *punane pliiats ja roheline pliiats* on sõnavorme kokku neli. Sõnavara puhul on oluline rõhutada, et selles töös räägitakse sõnavarast kui leksikonist, mida hinnatakse sõnavormide alusel. Oletame, et lapse repertuaaris on vaid üks lause *võtan punase pliiatsi ja rohelise pliiatsi*, siis selle lapse sõnavara suurus on 5 sõna.

CHILDES-i andmebaasis on eesti laste suulise kõne lindistused olnud alates 1998. aastast. 2016. aasta märtsikuu seisuga koosneb eesti lapsekeele korpus seitsmest alamkorpusest, mis on oma nimed saanud korpuse koostajate järgi: Argus, Beek, Kapanen, Kohler, Kõrgesaar, Vija ja Zupping (CHILDES, 2016). Arguse, Beeki, Kapaneni, Vija ja Zuppingu korpustes on andmete kogumiseks kasutatud longituuduurimusele omast viisi, kus teatud perioodi jooksul salvestatakse lapse ja hoidja vestlusi intensiivselt ja mille tulemusena tekib tihe andmestik. Kohleri ja Kõrgesaare korpustes pärinevad andmed salvestustest mitme eri vanuses lapsega. Lisades 1–6 on iga alamkorpuse kohta koondlikult välja toodud lapse nimi, vanus, sugu, sessioonide arv, lapse ja hoidja sõnade koguarv igas vanuses.

Kõikides alamkorpustes on alamkorpuse sõnade koguarvu poolest ülekaalus hoidja sõnad, mis on küllatki ootuspärane tulemus (vt joonis 2).



Joonis 2: hoidja ja lapse sõnade jaotumine alamkorpustes

Kõige mahukamad on Vija, Kõrgesaare ja Kohleri alamkorpused. Neist mahukaim on Vija korpus (vt joonis 2 ja lisa 1), sisaldades 226311 sõna, millest 49% (110381 sõna) olid lapse sõnad ja hoidja sõnad 51% (115930 sõna). Lindistusi tehti Andreasega vahemikus 1;7–3;1 eluaastat. Kõrgesaare korpus (vt ka lisa 5) koosneb 159511 sõnast, neist 41% (65819) on lapse sõnad ja 59% (93692) hoidja sõnad. Materjal pärineb lindistustest 12 erineva lapsega vahemikus 1;3–14;1 eluaastat. Siia pole sisse arvestatud transkriptsioone vestlustest, mille osalejateks olid vaid täiskasvanud. Kohleri korpus (vt ka lisa 7) sisaldab 70712 sõna, kus lapse sõnad moodustavad 14% (10053 sõna) ja hoidja sõnad 86% (60659 sõna). Lindistusi tehti 8 erineva lapsega vahemikus 0;11–2;3 eluaastat.

Beeki korpus (vt ka lisa 3) sisaldab 59136 sõna, millest lapse sõnad moodustavad 17% (10164 sõna) ja hoidja sõnad 83% (48972). Lindistusi tehti Liisbetiga vahemikus 0;9–2;5 eluaastat. Kapaneni korpus (vt ka lisa 4) koosneb 45675 sõnast, neist 35% (15938 sõna) on lapse sõnad ja 65% (29737) hoidja sõnad. Kapaneni materjal pärineb lindistustest Martinaga vahemikus 1;3–2;7 eluaastat. Zuppingu korpus (vt ka lisa 6) sisaldab 42461 sõna, neist 27% (11632) on lapse sõnad ja 73% (30829) hoidja sõnad. Kõik lindistused on tehtud Lindaga vahemikus 1;3–4;2 eluaastat. Mahult kõige väiksem on Arguse korpus (vt ka lisa 2). Korpus sisaldab 13373 sõna, millest 32% (4220) on lapse ja 68% (9153) hoidja sõnad. Lindistusi

tehti Hendrikuga vahemikus 1;8–2;5 eluaastat.

Tabel 1 näitab, kuidas jaotuvad kogu korpuse lapse ja hoidja sõnad alamkorpuste kaupa.

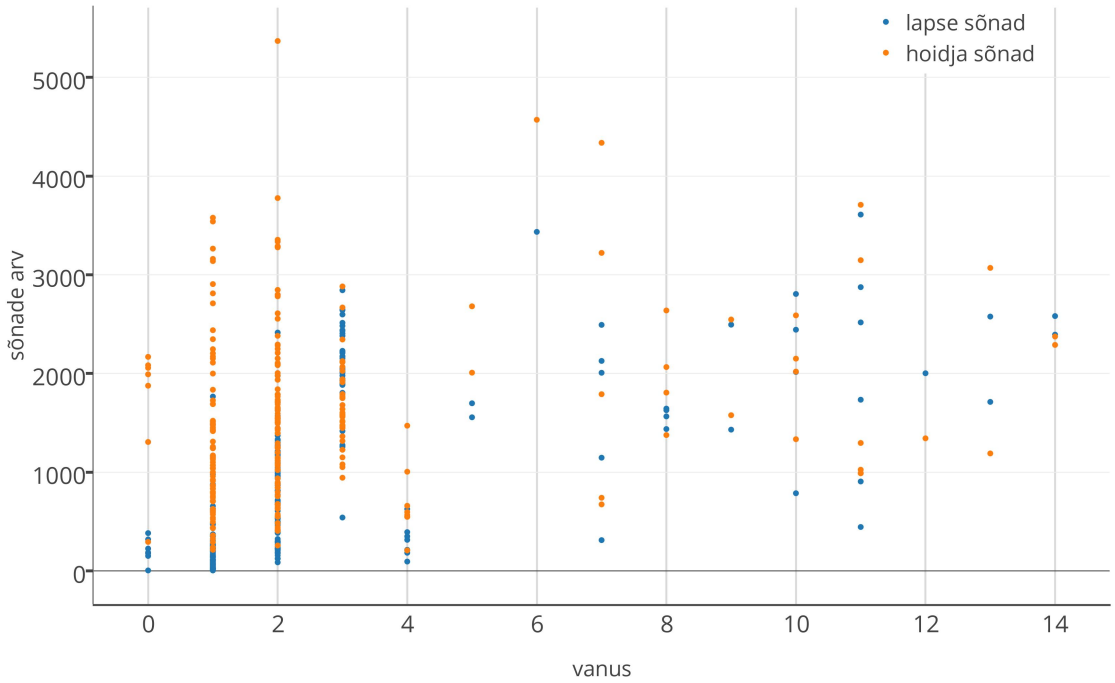
korpus	lapse sõnad	% kogu korpuses	hoidja sõnad	% kogu korpuses
Vija	110381	48%	115930	30%
Kõrgesaar	65819	29%	93692	24%
Argus	4220	2%	9153	2%
Beek	10164	4%	48972	13%
Kapanen	15938	7%	29737	8%
Zupping	11632	5%	30829	8%
Kohler	10053	4%	60659	16%
KOKKU	228207 37%	100%	388972 63%	100%

Tabel 1: lapse ja hoidja sõnade % kogu korpuses

Kogu korpuses moodustavad lapse sõnad 37% (228207 sõna) ja hoidja sõnad 63% (388972 sõna). Korpuse suuruseks on 617179 sõna. Kogu korpuses on Vija alamkorpus nii lapse kui hoidja sõnade poolest kõige suurema osakaaluga: lapse sõnad moodustavad 48% kõigi laste sõnade arvust, hoidja sõnad 30% kõigi hoidja sõnade arvust (vt tabel 1). Kõrgesaare alamkorpus moodustab kõigi laste sõnade arvust 29%, hoidja sõnad 24% kõigi hoidja sõnade arvust. Kohleri alamkorpuse lapse sõnad moodustavad 4% kõigi laste sõnade arvust, hoidja sõnad 16% kõigi hoidja sõnade arvust. Beeki alamkorpuse lapse sõnad moodustavad 4% kõigi laste sõnade arvust, hoidja sõnad 13% kõigi hoidja sõnade arvust.

Kapaneni alamkorpuse lapse sõnad moodustavad 7% kõigi laste sõnade arvust, hoidja sõnad 8% kõigi hoidja sõnade arvust. Zuppingu alamkorpuse lapse sõnad moodustavad 5% kõigi laste sõnade arvust, hoidja sõnad 8% kõigi hoidja sõnade arvust. Arguse alamkorpuse lapse sõnad moodustavad 2% kõigi laste sõnade arvust, hoidja sõnad samuti 2% kõigi hoidja sõnade arvust.

Korpuse üks koostamise põhimõte on, et korpus peab olema representatiivne ehk uuritava nähtuse suhtes esinduslik. Kui me uurime nii lapse kui hoidja keelekasutust, siis tuleb teada, kelle keelekasutust korpus esindab. Joonis 3 illustreerib seda, kuidas jaotuvad lapse ja hoidja sõnad kogu korpuses. Me näeme, et suurem osa transkriptsioonidest on tehtud vanuses 1 kuni 3. See tähendab, et korpuses on suur hulk sõnu, mis esindavad vaid teatud vanusegruppe ja see omakorda peegeldab paljude teiste vanusegruppide esinduslikkust.

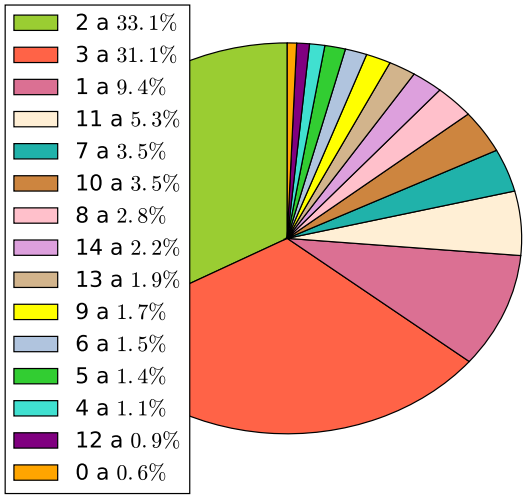


Joonis 3: hoidja ja lapse sõnade jaotumine kogu korpuses vanuse järgi

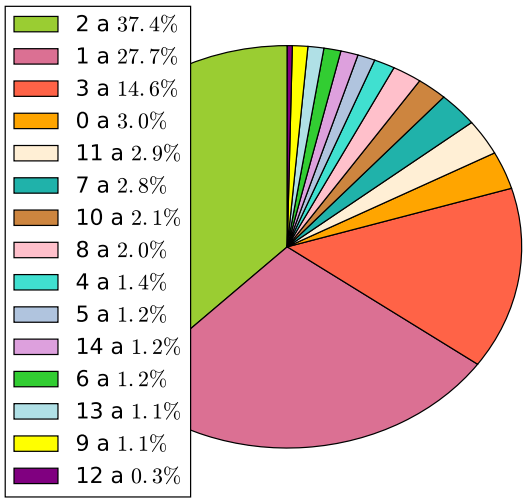
Kui vaadata detailsemalt lapse ja hoidja sõnade jaotumist vanusegrupiti, siis kogu korpus esindab 2- ja 3-aastaste laste keelekasutust (vt joonis 4 (a)). Ülejäänud vanusegrupid (ehk 86% kõikidest vanusegruppidest) esindavad korpuses lapse sõnu vaid 35,8%. Hoidja sõnadest (vt joonis 4 (b)) on kõige suurema osakaaluga 2-aastaste (37,4%), 1-aastaste (27,7%) ja 3-aastaste (14,6%) vanusegrupp. Ülejäänud vanusegrupid (ehk 80%) esindavad hoidjate sõnadest vaid 20,3%. Selline tulemus on loogiliselt põhjendatav: mida rohkem sessioone, seda rohkem sõnu ja seda suurem korpus. Kõigis alamkorpustes tehti kõige enam lindistusi just 1-, 2- ja 3-aastaste lastega.

Korpus esindab enam-jaolt Andrease (Vija korpus) ja tema hoidja keelekasutust, sest Andrease sõnad moodustavad kõikide laste sõnadest lausa 48,4% ja Andrease hoidja sõnad kõikidest hoidja sõnadest 29,8% (vt joonis 5 (a) ja (b)). See tähendab, et ülejäänud lapsed (ehk 96%) esindavad kogu korpuses lapse keelekasutust vaid natuke üle 50%. Ja nendest ülejäänud lastest 56% (ehk 14 last 25 lapsest) esindavad korpuses vähem kui 1%. Hoidjate keelekasutus varieerub natuke rohkem: kogu korpuses esindab ülejäänud hoidjate keelekasutus 70,2%, kuid hoidjatest 32% (ehk 8 hoidjat 25 hoidjast) esindab kogu korpuses vähem kui 1% hoidja sõnu. Ka siin mängivad olulist rolli lapsega tehtud sessioonide arv ja korpuse suurus. Näiteks Vija korpuse puhul on tegemist tiheda andmestikuga, kus Andreasega tehti kokku

74 sessiooni, kuid Kõrgesaare korpuses tehti Sirlini, Arturi ja Arabellaga vaid 1 sessioon, mistõttu on kogu korpuses nende laste keelekasutuse osakaal miniatuurne.

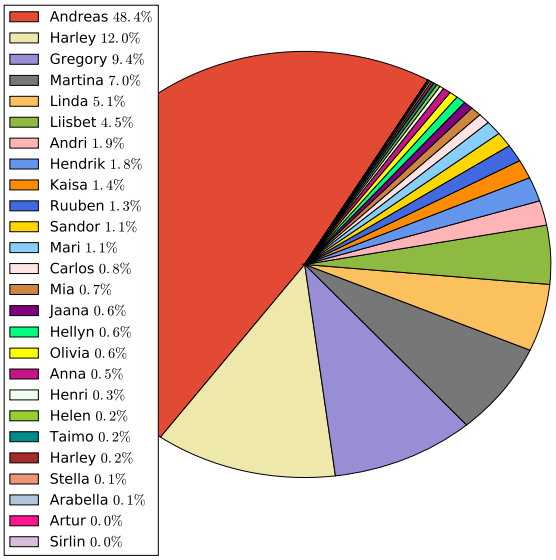


(a) lapse sõnad

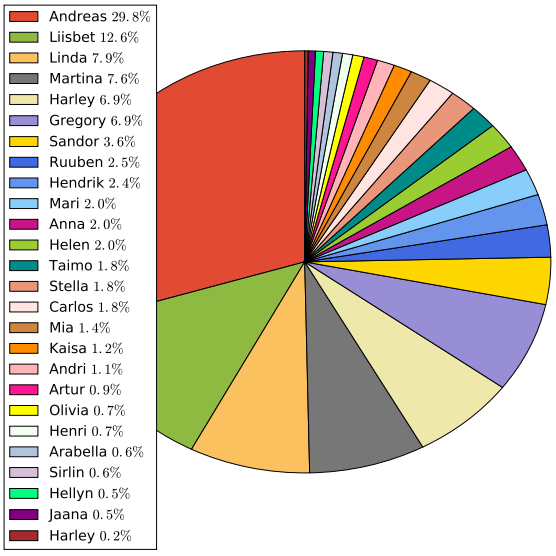


(b) hoidja sõnad

Joonis 4: lapse ja hoidja sõnade jaotumine kogu korpuses vanusegrupiti



(a) lapse sõnad



(b) hoidja sõnad

Joonis 5: lapse ja hoidja sõnade jaotumine lapse järgi

Tabelis 2 on välja toodud see, kuidas jaotub lapse, hoidja ja nende kattuv sõnavara (ehk unikaalsete sõnade koguarv) igas alamkorpuses. Kõige ühtlasemalt jaotub lapse ja hoidja sõnavara Kapaneni korpuses (mõlema osakaal korpuses on 39%) ja nende kattuv sõnavara on 22%. Ka Kõrgesaare korpuses on jaotumine ühtlane – lapse sõnavara moodustab 35% ja hoidja sõnavara 36%, seejuures kattuv sõnavara on 29%. Vija ja Zuppingu korpusteski pole lapse ega hoidja sõnavaral suured erinevused (35% ja 32% ning 29% ja 54%). Kõige suuremad lahknevused on Beeki, Arguse ja Kohleri korpuses. Kohleri korpuses moodustab lapse sõnavara suurus kogu alamkorpuse sõnavarast 17% ja hoidja sõnavara 58%. Arguse korpuses moodustab lapse sõnavara kogu alamkorpuse sõnavarast 17% ja hoidja sõnavara 69%. Sõnavara jaotub kõige ebaühtlasemalt Beeki korpuses: lapse sõnavara moodustab 16% ja hoidja sõnavara 78%, nende ühine sõnavara alamkorpuse kogu sõnavarast on vaid 6%. Kõige suurem ühise sõnavara % on Vija alamkorpuses (33%). Vija korpuse puhul on huvitav see, et see on ainus alamkorpus, kus lapse sõnavara on rikkam kui hoidja sõnavara.

korpus	lapse sõnavara	hoidja sõnavara	kattuv sõnavara	KOKKU
Argus	291 17%	1185 69%	248 14%	1724 100%
Beek	732 16%	3517 78%	272 6%	4521 100%
Kapanen	2661 39%	2708 39%	1533 22%	6902 100%
Kohler	817 17%	2700 58%	1160 25%	4677 100%
Kõrgesaar	5402 35%	5665 36%	4455 29%	15522 100%
Vija	4834 35%	4484 32%	4643 33%	13961 100%
Zupping	1615 29%	3016 54%	982 17%	5613 100%

Tabel 2: lapse ja hoidja sõnavara jaotumine alamkorpustes

Sõnavara kvantitatiivseks hindamiseks tuleb arvestada korpuse andmete kogumisviisi ja suurusega. Väikse andmestikuga korpused pole sõnavara ja harva esilduvate nähtuste uurimiseks kõige sobilikum viis, sest see võib viia valede järeldusteni. Tihe andmestik võimaldab jällegi saada rikkalikumat sõnavara, uurida lapse produktiivsust ja nähtusi, mis pole keeles nii sagedased. (Tomasello & Stahl, 2004) Kui vaadata Beeki, Arguse ja Zuppingu alamkorpustes lapse ja hoidja sõnavara

kattumist, siis tekib küsimus, kas see vähene kattuvus tuleneb sellest, et lapsel ongi väike sõnavara, või sellest, et tegemist väikse korpusega? Alamkorpustes ei saa vanuse järgi lapse ja hoidja sõnavara kasvu adekvaatselt hinnata, sest tegu on väikeste andmestikega ja mõnes vanuses on last rohkem lindistatud. Lisaks tuleks ettevaatlikkusega suhtuda ka alamkorpuste võrdlemisse, sest iga korpus on koostatud erinevaid eesmärke ja viise silmas pidades.

4. Morfoloogiliselt märgendatud lapsekeele korpus

4.1. Tööprotsess

Magistritöö eesmärk on luua morfoloogiliselt märgendatud eesti lapsekeele korpus, kuhu on koondatud kõik CHILDES-i eesti keele alamkorpused. Esialgne plaan oli konverteerida omalkäel kõik CHAT-failid XML-kujule, kuid sellega tekkisid mõningad tagasilöögid. Selleks, et faile XML-kujule konverteerida, oleks tarvis, et kõik alamkorpused oleksid ühtsel kujul transkribeeritud ja kodeeritud. Peatükis 3.2 tõin välja mõned näited sellest, kui ebajärjepidevalt on seda tehtud. Isegi, kui korpused oleksid olnud standardsel kujul, siis oleks konverteerimisskripti tegemine muutunud väga keeruliseks ja ülejõukäivaks ülesandeks. Põhjus seisneb selles, et CHILDES-i transkriptsioonisüsteemis on väga suur ja lai valik kodeeringuid, mida on paraku ühel inimesel raske hallata. Näide (5) illustreerib seda, kuidas juba ühes lühikeses transkriptsiooni lõigus võib kodeeringute kasutus olla väga mitmekesine (kodeeringu seletus paikneb lausungi järel).

(5)

*CHI: see kifiir **[:** kefiir**]** . | *asendus*

*MOT: kus sa **+/**. | *vahele segamine*

*CHI: **+ <** (h)akkas põlema . | *pealeräägimine*, *mittetäielik sõna*

*MOT: see ei ole kefiir ju .

*CHI: kefiir . **[+** sr**]** | üleskirjutaja poolt defineeritud *postcode*

*MOT: see on piim .

*FAT: mis see kook teeb ?

*FAT: tuleb ära panna **[=** visata**]** või ? | *seletus, tähendus*

*MOT: mina ei tea , vist jah .

*CHI: kuidas emme küpsetab saia , lihat**@n [*]** . | *üleüldistamine*, *vea markeerimine*

%err: lihat=liha $MOR

*FAT: saia ei ei küpseta .

*FAT: kartulit küpsetame , **(.)** ahjus . | *paus*

*CHI: **+ <** saia . | *pealerääkimine*

*CHI: lihat**@n [*]** . **[+ sr]** | *postcode*

%err: lihat=liha $MOR

*FAT: liha ka jah .

*CHI: lihat**@n [*]** . **[+ sr]** | *üleüldistamine*, *vea markeerimine*, *postcode*

%err: lihat=liha $MOR

%act: MOT koorib sibulat

...

*CHI: Atu **[:** Andreas**]** sõi **+...** | *asendus*, *kõrvalekalle*

*CHI: **+"** a . | *lausung jutumärkides*

...

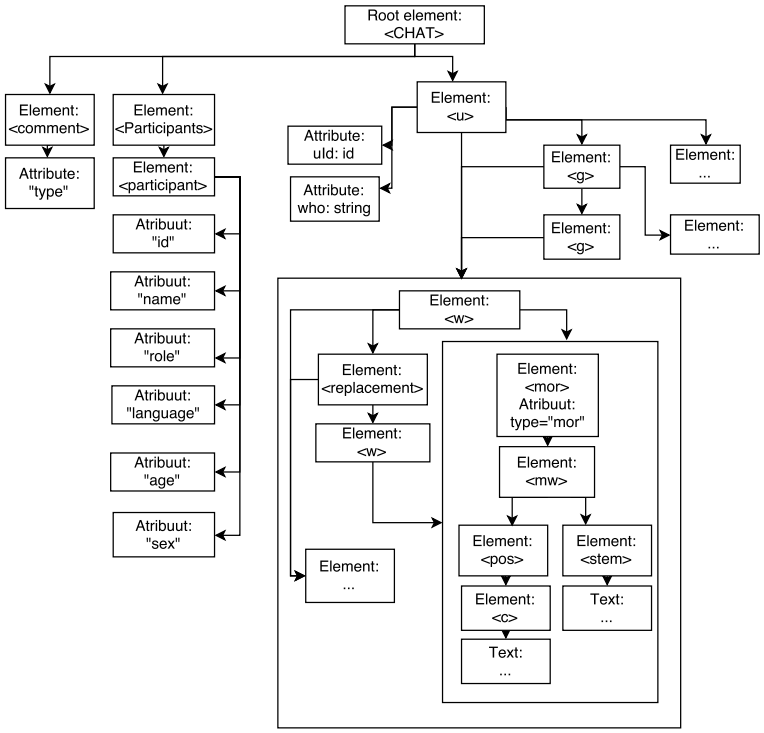
*CHI: käpad (h)aige **[*]** **[/]** käpad (h)aige **[*]** . | *mittetäielik sõna*, *vea markeerimine*, *kordus*

(Vija; 20018.cha)

On arusaadav, et iga uurija transkribeerib ja kodeerib lindistusi lähtuvalt enda eesmärkidest. Ühelt poolt on hea, et transkriptsioonisüsteem on niivõrd detailne, kuid teisalt võib selles orienteerumine vägagi raskeks osutuda. Kuna sellise konverteerimisskripti kirjutamise töömaht oleks selle magistritöö kirjutamise jaoks liiga töömahukaks osutunud, siis tuli leida uus lahendus. Korpuse tegemiseks vajalik keelematerjal pärineb samuti CHILDES-i andmebaasist, kuid need on juba eelnevalt CHAT-kujult XML-kujule konverteeritud failid. Aga kuna selle töö eesmärk on luua morfoloogiliselt märgendatud korpus, siis tuli nendele XML-failidele ka lisada morfoloogiline tasand, mida neis failides ei ole. Kuna kõik CHILDES-i korpused on kirjeldatud oma XML-skeema (*Talkbank*) järgi, siis tuli lähemalt tutvuda Talkbanki XML-skeema süntaksiga.

4.1.1. *Talkbank*i skeema

Joonisel 6 on kujutatud minu töö seisukohalt olulisimad skeema elemendid.



Joonis 6: Talkbanki skeema elemendid

Talkbanki skeema juurelement on *<CHAT>*, mille alluvad on *<comment>*, *<Participants>* ja *<u>*. Elemendi *<Participants>* alluv on *<participant>*. *<participant>* elemendis peitub metainfo lindistuse osalejate kohta (kõneleja ID, nimi, roll, keel, vanus ja sugu). Element *<comment>* talletab metainfot lindistuse konteksti kohta (nt koht, kuupäev, lindistuse algus ja lõpp jne).

Näide (6) (Argus; hend10.xml)

```
<Participants>
  <participant
    id="CHI"
    name="Hendrik"
    role="Target_Child"
    language="est"
    age="P2Y2M6D"/>
```

```
    <participant
      id="EMA"
      role="Mother"
      language="est"/>
  </Participants>
  <comment type="Date">02-JUN-1997</comment>
```

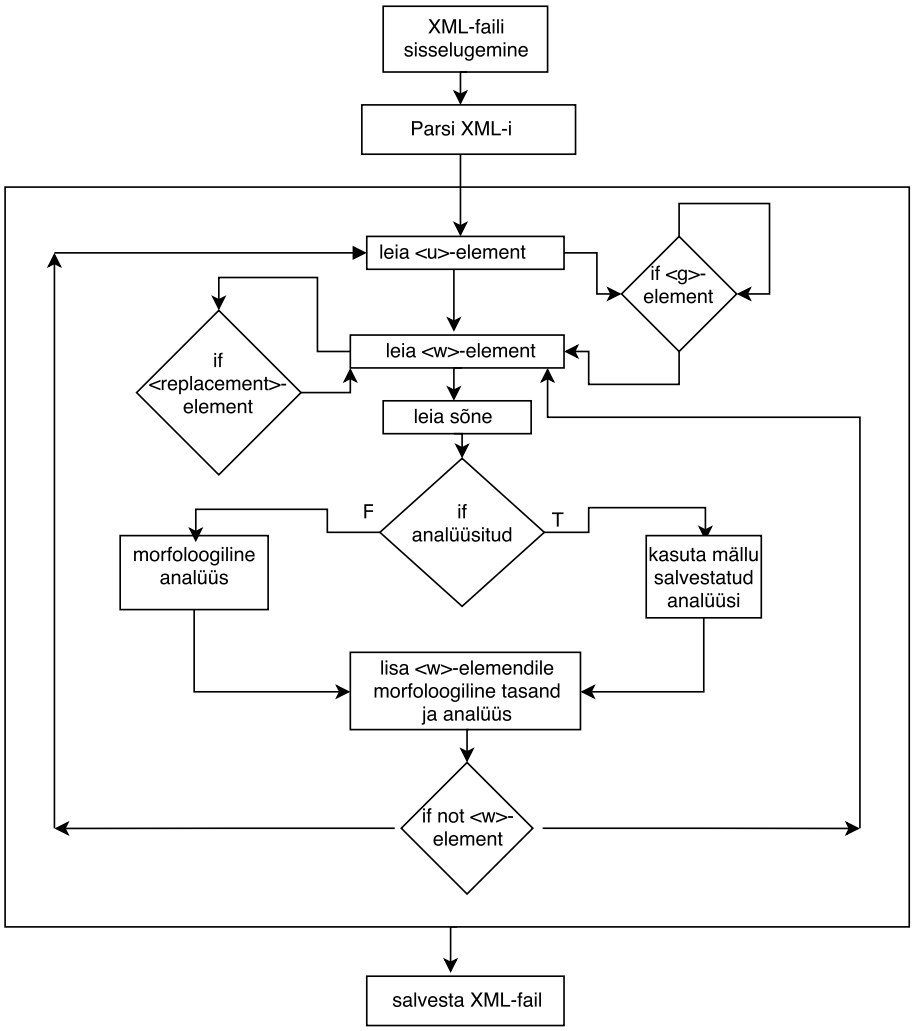
Element *<u>* tähistab kõneleja lausungit, selle kohustuslikeks atribuutideks on kõneleja ID ja lausungi järjekorra ID. *<u>*-elemendil on palju alluvaid, aga selle töö juures osutusid olulisimateks *<w>* ja *<g>*. Element *<w>* tähistab sõna ja *<g>* sõnade gruppi. *<g>* alluvaks võib olla tema ise või *<w>*. Elemendi *<w>* alluvaks võib olla *<replacement>*. See element tähistab neid üksuseid, mida CHILDES-i konventsioonide järgi kodeeritakse [: text] abil, vt näide (7) (vt ka näide (5) kifiir [: kefiir]).

Näide (7) (Kohler; car030900.xml)

```
<u who='CHI' uID='u58'>
  <w>ehitame</w>
  <w>
      galaasi
      <replacement>
          <w>garaazhi</w>
      </replacement>
  </w>
  <t type='p'></t>
</u>
```

Morfoloogilise tasandi lisamine algab elemendiga *<mor>*, mille alluv on *<mw>* ehk *morphemicWordType*. See jaguneb elemendiks *<pos>* ja *<stem>*. *<pos>* tähistab sõnaliiki (ingl k. *part of speech*). Selle alluvaks on *<c>* ehk sõna morfoloogiline kategooria, mille sisuks on mittetühi string. Element *<stem>* tähistab sõnatüve, mille sisuks on samuti mittetühi string.

4.1.2. Morfoloogilise info lisamine



Joonis 7: Töövoog

Programmi kirjutamiseks on kasutatud Python 3.4 versiooni. Töövoog (vt joonis 7) on jaotatud 4 suuremaks osaks: XML-failide sisselugemine, faili parsimine, töötlemine ja modifitseeritud XML-faili salvestamine. Faili parsimiseks kasutan Pythoni moodulit *ElementTree*, mis võimaldab lugeda ja genereerida XML hierarhiat. Parsimise käigus pannakse paika XML-faili juurelement ja sellele alluvad elemendid. Töötluse käigus navigeeritakse esmalt iga vestlusest osavõtja lausungi juurde. Seejärel leitakse kõik sõned ehk *<w>*-elemendid. Juhul kui *<w>*-elemendi alluv on element *<replacement>*, siis uueks sõneks määratakse *<replacement>*-elemendi

<*w*>-element.

Kui lausungi sõne on leitud, siis tehakse sõnele morfoloogiline analüüs. Morfoloogilise analüsaatorina kasutatakse *etanat*. Sõne analüüs salvestatakse mällu. Kui analüsaator saab sisendiks seni nägemata sõne (ehk mida mälus ei eksisteeri), siis tehakse sellele morfoloogiline analüüs. Põhjus seisneb programmi optimeerimises: morfoloogilise analüsaatori kutsumine iga sõne juures on üsna kulukas protsess. Seejärel lisatakse igale sõnele morfoloogiline tasand. Programm genereerib kirjutatud koodis järk-järgult puu elemendid ning morfoloogilisele tasandile jõudes hakkab sõne analüüse neile vastavatesse elementidesse lisama (vt joonis 7 ja näide (8)). Juhul kui sõne analüüse on rohkem kui üks, siis igale analüüsile genereeritakse uuesti morfoloogilise taseme elemendid. Samme korratakse seni, kuni jõutakse viimase lausungini ja lõpptulemus salvestatakse modifitseeritud XML-faili.

Näide (8) (Vija; 11120.xml)

```
<u uID="u7" who="CHI">
  <w>tuli
      <mor type="mor">
          <mw>
              <pos><c>_V_ Pers Prt Ind Sg3 Aff</c></pos>
              <stem>tule+i</stem>
          </mw>
      </mor>
      <mor type="mor">
          <mw>
              <pos><c>_S_ Sg Nom</c></pos>
              <stem>tuli+0</stem>
          </mw>
      </mor>
  </w>
   ...
```

4.2. Morfoloogilise märgenduse hindamine

Analüüsisin korpuseid morfoloogilist analüsaatorit kohandamata. Analüüsimisel ei teostatud oletamist ega ühestamist. Selle alapeatüki eesmärk on anda ülevaade sellest, kuidas jaotuvad analüüsi saanud ja tundmatuks jäänud sõnad igas vanuserühmas alamkorpuse kaupa.

vanus	lapse sõnad			hoidja sõnad		
	analüüsitud	tundmatud	KOKKU	analüüsitud	tundmatud	KOKKU
1	515 47%	575 53%	1090 100%	11033 93%	838 7%	11871 100%
2	3208 80%	818 20%	4026 100%	10030 95%	534 5%	10564 100%
3	2292 90%	242 10%	2534 100%	5232 94%	319 6%	5551 100%
4	1439 81%	330 19%	1769 100%	3929 95%	210 5%	4139 100%
5	2947 91%	309 9%	3256 100%	4531 97%	159 3%	4690 100%
6	3222 94%	213 6%	3435 100%	4435 97%	135 3%	4570 100%
7	7572 94%	518 6%	8090 100%	10489 97%	278 3%	10767 100%
8	5878 94%	400 6%	6278 100%	7522 95%	367 5%	7889 100%
9	3658 93%	269 7%	3927 100%	3879 94%	246 6%	4125 100%
10	7536 94%	518 6%	8054 100%	7717 95%	378 5%	8095 100%
11	11240 93%	851 7%	12091 100%	10762 96%	399 4%	11161 100%
12	1847 92%	156 8%	2003 100%	1274 95%	70 5%	1344 100%
13	3798 89%	492 11%	4290 100%	4062 95%	199 5%	4261 100%
14	4525 91%	451 9%	4976 100%	4318 93%	347 7%	4665 100%
KOKKU	59677 91%	6142 9%	65819 100%	89213 95%	4479 5%	93692 100%

Tabel 3: Kõrgesaare korpus

vanus	lapse sõnad			hoidja sõnad		
	analüüsitud	tundmatud	KOKKU	analüüsitud	tundmatud	KOKKU
1	5093 70%	2209 30%	7302 100%	16584 93%	1207 7%	17791 100%
2	5689 83%	1142 17%	6831 100%	9248 94%	583 6%	9831 100%
3	1601 89%	204 11%	1805 100%	2067 98%	48 2%	2115 100%
KOKKU	12383 78%	3555 22%	15938 100%	27899 94%	1838 6%	29737 100%

Tabel 4: Kapaneni korpus

vanus	lapse sõnad			hoidja sõnad		
	analüüsitud	tundmatud	KOKKU	analüüsitud	tundmatud	KOKKU
0	455 31%	995 69%	1450 100%	10282 90%	1205 10%	11487 100%
1	297 26%	846 74%	1143 100%	9605 92%	858 8%	10463 100%
2	5034 66%	2537 34%	7571 100%	25196 93%	1826 7%	27022 100%
KOKKU	5786 57%	4378 43%	10164 100%	45083 92%	3889 8%	48972 100%

Tabel 5: Beeki korpus

vanus	lapse sõnad			hoidja sõnad		
	analüüsitud	tundmatud	KOKKU	analüüsitud	tundmatud	KOKKU
0	6 100%	0 0%	6 100%	281 95%	14 5%	295 100%
1	4592 93%	348 7%	4940 100%	40724 97%	1160 3%	41884 100%
2	4992 98%	115 2%	5107 100%	18224 99%	256 1%	18480 100%
KOKKU	9590 95%	463 5%	10053 100%	59229 98%	1430 2%	60659 100%

Tabel 6: Kohleri korpus

vanus	lapse sõnad			hoidja sõnad		
	analüüsitud	tundmatud	KOKKU	analüüsitud	tundmatud	KOKKU
1	2213 78%	632 22%	2845 100%	8391 98%	130 2%	8521 100%
2	39313 95%	2185 5%	41498 100%	57989 98%	1283 2%	59272 100%
3	64231 97%	1807 3%	66038 100%	47466 99%	671 1%	48137 100%
KOKKU	105757 96%	4624 4%	110381 100%	113846 98%	2084 2%	115930 100%

Tabel 7: Vija korpus

vanus	lapse sõnad			hoidja sõnad		
	analüüsitud	tundmatud	KOKKU	analüüsitud	tundmatud	KOKKU
1	361 64%	205 36%	566 100%	1884 96%	79 4%	1963 100%
2	3092 85%	562 15%	3654 100%	6945 97%	245 3%	7190 100%
KOKKU	3453 82%	767 18%	4220 100%	8829 96%	324 4%	9153 100%

Tabel 8: Arguse korpus

vanus	lapse sõnad			hoidja sõnad		
	analüüsitud	tundmatud	KOKKU	analüüsitud	tundmatud	KOKKU
1	2508 68%	1169 32%	3677 100%	14756 97%	435 3%	15191 100%
2	5782 85%	1003 15%	6785 100%	12882 98%	232 2%	13114 100%
3	470 87%	72 13%	542 100%	1028 98%	24 2%	1052 100%
4	596 95%	32 5%	628 100%	1455 99%	17 1%	1472 100%
KOKKU	9356 80%	2276 20%	11632 100%	30121 98%	708 2%	30829 100%

Tabel 9: Zuppingu korpus

Võtame esmalt vaatluse alla hoidja sõnad. Tabelitest 3–9 on näha, et nii analüüsi saanud kui ka tundmatuks jäänud sõnade osakaal igas alamkorpuses on üsna

stabiilselt jaotunud. Analüüsitud sõnade üldine osakaal varieerub alamkorpusetes 94%–98% vahel ja tundmatute sõnade osakaal 2%–8% vahel. 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 korpuste esmasel morfoloogilisel analüüsimisel saadi tundmatute sõnade protsendiks jututubades 27,2%, foorumites 10.3%, kommentaariumites 5,6% ja uudisgruppides 11,7% (Muischnek et al., 2016). Pärast kasutajasõnastiku ja eeltöötluse rakendamist vähenes tundmatute sõnade protsent jututubades 10,5%, foorumites 8.8%, kommentaariumites 4,8% ja uudisgruppides 10,5%-ni. Seega väike tundmatute sõnade % hoidja keeles on hea, kuid need andmed võivad viidata ka sellele, et korpuse transkribeerijad on pannud hoidjakeelt kirja kirjakeelele sarnaselt.

Lapse sõnade puhul varieerub analüüsitud sõnade üldine osakaal igas alamkorpuses 57%–96% vahel ja tundmatud sõnad 4%–43% vahel. Kõige suurem tundmatute sõnade osakaal on Beeki alamkorpuses (vt tabel 5). Seal varieerub tundmatute sõnade % 34 ja 74 vahel. Beekile järgneb Kapaneni alamkorpus (vt tabel 4), kus tundmatute sõnade % varieerub 11 ja 30 vahel. Zuppingu alamkorpuses (vt tabel 9) varieerub 5 ja 32% vahel ning Arguse alamkorpuses (vt tabel 8) 15 ja 36% vahel. Kõige väiksem tundmatute sõnade osakaal on Vija (vt tabel 7) ja Kohleri alamkorpuses (vt tabel 6). Vija korpuses varieeruvad tundmatud sõnad 3 ja 22 % vahel, kuid Kohleri korpuses 0–7 % vahel. Kõrgesaare korpuses (vt tabel 3) varieerub tundmatute sõnade osakaal vahemikus 6%–53%. Olenemata sellest, et tundmatute sõnade % amplituut on suur, jääb üldine skoor alla 10%. Huvitav see, et kui teiste korpuste puhul üldjuhul vanuse suurenemisega kahaneb tundmatute sõnade osakaal, siis näiteks Kohleri korpuses 1. vanusegrupp ehk 0-aastaste puhul on tundmatuid sõnu 0%. Sellesse peab natuke kriitiliselt suhtuma, sest see tuleneb väiksest sõnade koguarvust. Sama on ka Beeki korpuses, kus 0-vanusegrupi protsent on väiksem kui 1-aastaste seas (69% vs 74%).

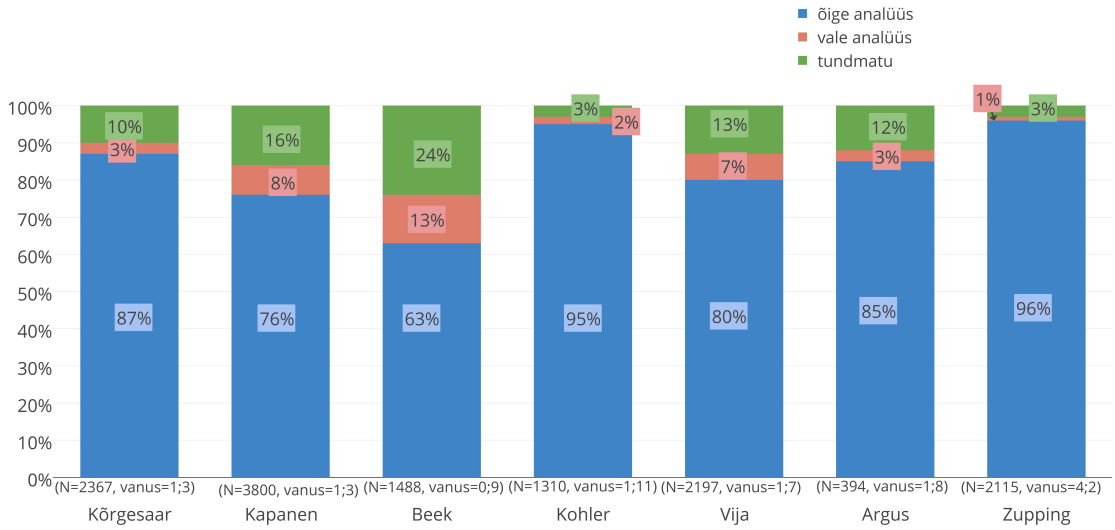
Need andmed võivad viidata sellele, et tundmatute sõnade osakaal võib sõltuda paljuski transkribeerijast, täpsemalt transkribeerija üleskirjutamise viisist. Alamkorpuste standardiseerimise alapeatükis kirjutasin, et transkribeerija peab nägema ja teadma, mida lindistuses tegelikult öelda taheti, ja vastavad kodeeringud peaksid transkriptsioonis olema nii lisatud, et vead oleksid juba esimesel tasandil liigitatud. Selleks, et morfoloogiline analüsaator saaks oma tööd hästi teha, oleks tarvis, et transkribeerijad kasutaksid üleskirjutamisel kodeeringut ([: sõna]), mis asendab kõneleja poolt produtseeritud sõna selle kirjakeelele vastava sõnaga. Näiteks vaatame Kohleri korpuses 1. vanusegruppi, kus tundmatute sõnade osakaal

on 0%. Lapse sõnu on kokku 6 ja sõnavorme 4:

lapse sõna	asendus	analüüs
teh	tere	_I_ tere+0
tehe	tere	_I_ tere+0
täh	aitäh	_I_ aitäh+0
äh	aitäh	_I_ aitäh+0

Me näeme, et kontekstita need sõnad justkui ei tähenda miskit, aga kuna nendele sõnadele on juurde lisatud kodeering selle kohta, mida need tegelikkuses tähendavad või mida laps üritas öelda, siis nii saab morfoloogiline analüsaaator oma tööga hästi hakkama ja seetõttu pole selle lapse keelekasutuses ühtki tundmatut sõna. See annab alust arvata, et neis alamkorpustes, kus tundmatute sõnade osakaal on väike, kasutatakse kõne transkribeerimisel kodeeringut [: sõna], nt Zuppingu korpuses pole seda kordagi kasutatud.

Seni olen arutlenud vaid tundmatute sõnade teemal, kuid morfoloogilise analüsaatori adekvaatsuse hindamiseks tuleb vaatluse alla võtta ka analüüsi saanud sõnad. Selleks oli tarvis faile käsitsi läbi vaadata. Igast korpusest valisin juhuslikkuse alusel ühe faili ja hindasin iga sõna puhul, kas saadud analüüs on õige või mitte. Joonisel 8 on kujutatud, kuidas jaotuvad alamkorpustes tundmatud, õige ja vale analüüsi saanud sõnad.



Joonis 8: tundmatud, õige ja vale analüüsi saanud sõnad alamkorpuse kaupa

Kõige paremad tulemused olid Zuppingu alamkorpuses, kus keelematerjal pärineb lindistusest lapsega vanuses 4;2. Kõikidest sõnadest oli vale analüüsi saanud sõnu vaid 1%, õige analüüsi saanud sõnu 96% ja tundmatuid sõnu 3%. Kohleri alamkorpuses (lapse vanus 1;11) olid kõikidest sõnadest vale analüüsi saanud 2%, õige analüüsi 95% ja tundmatuks jäänud sõnu 3%. Nii Arguse kui ka Kõrgesaare korpuses olid kõikidest sõnadest 3% saanud vale analüüsi (Argusel laps vanuses 1;8 ja Kõrgesaarel 1;3). Vija alamkorpuses (laps vanuses 1;7) oli kõikidest sõnadest vale analüüsi saanud 7%. Kapaneni korpuses (laps vanuses 1;3) oli 8% sõnadest saanud vale analüüsi. Protsentuaalselt oli kõige enam vale analüüsi saanud sõnu (13%) Beeki korpuses (laps vanuses 0;9).

Siinkohal tuleks muidugi tähelepanu pöörata ka sellele, et tekkinud on tahtmatu vanuseline järjestus – kõige vähem vale analüüsi saanud sõnu on Zuppingu korpuses, kus lapse vanus on 4;2, ja kõige enam Beeki korpuses, kus lapse vanus on 0;9. Paraku pole vale analüüsi saanud sõnadel vanusega midagi pistmist, pigem on küsimus transkriptsioonide üleskirjutajas. Kui me vaatame Kapaneni ja Beeki alamkorpustes (vt tabel 4 ja 5) tundmatute ja vale analüüsi saanud sõnade osakaalu, siis näeme, et just neis korpustes on need kõige suuremad ja just nendes korpustes kasutatakse kõige enam nii lapse kui hoidja kõne transkribeerimisel kuuldeortograafiat ja sedagi mitte järjepidevalt. Näiteks Kapaneni ja Beeki korpuses kasutatakse läbisegi *head isu* ja *ead isu*, *präegu* ja *praegu*, *jaaah* ja *jah*, *eiä* ja *ei*, *äitäh* ja *aitäh* jne. Lisaks veel sõnaalgulised klusiilid *buutuda* ja *puutuda*, *boti* ja *poti*, *balju* ja *palju* jne. Selline üleskirjutusviis mõjutab ka morfoloogilise analüsaatori väljundit nii tundmatute kui ka vale analüüsi saanud sõnade osas, nt sõna *ead* saab analüüsiks *iga+d _S_ Pl Nom iga+d*, sõna *präägust* saab analüüsiks *prääk+0 _S_ Sg Gen*.

Täielikult vigadeta morfoloogiliselt märgendatud korpus eeldab, et iga sõnavorm saab õige sõnaliigilise kuuluvuse, käändsõnade puhul õige arvu ja käände, verbide puhul õige arvu, isiku, tegumoe, aja, kõneviisi ja kõnelaadi. 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 käsitsi määramisel 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. (Muischnek & Vider, 2004, 102–104; Kaalep et al., 2000, 627–631).

Uue meedia korpuses võeti morfoloogilisel märgendamisel kasutusele partikli sõnaliik. Partikkel on muutumatu mittetäistähenduslik sõna, millel on eelkõige suhtluslik ja emotsionaalne funktsioon. (Muischnek et al., 2016, 4) Ka lapsekeele korpuse puhul tuleks mõelda mõne uue sõnaliigi kasutusele võtmise peale. Näiteks, mida teha onomatopoeetiliste sõnadega? 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; vt ka Laing, 2014b). Reili Argus eristab onomatopoeetiliste sõnade hulgas ka *imitatiive*, mis on onomatopoeetilised sõnad, mille häälikuline kuju võib olla varieeruv, kuid ei muutu morfoloogiliselt. Tüüpiliseks imitatiiviks on nt kiirabiauto signaali imiteeriv *viuviu*, kõndmise väljendamiseks kasutatav *tipa-tapa*. (Argus, 2004, 19–22)

Lisaks sellele, et onomatopoeetiliste sõnade ja imitatiivide piir on hägune, on onomatopoeetilised sõnad ja imitatiivid ka lapse varases keelekasutuses sõnaliigililt mitmesed. (Argus, 2004, 20–21) Eesti keele käsiraamat (EKK, 2007) jagab tähenduse järgi onomatopoeetilised sõnad interjektsioonide alla. Hennoste nimetab jällegi interjektsiooni sõnaliigiliseks prügikastiks, sest sinna on pandud kokku erinevad üksused. Hennoste arvates on onomatopoeetilised sõnad interjektioonide alla paigutatud sellepärast, et neil on kaldeline foneetiline ja fonoloogiline struktuur ning nad paiknevad sõna ja mittesõna piirimail. (Hennoste, 2002, 67) Näites (1) ja (2) jääb imitatiivi sõnaliigiline kuuluvus segaseks:

(1)

*CHI: **addrr**, **drrr**, **brrr**

*MOT: just, niimoodi sa õues sõidad vankriga (Argus, 2004, 27)

(2)

%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 enne presüntaktilist perioodi ongi raske sõnu liigitada ja sõnaliikidest saab alles siis rääkida,

kui laps hakkab kasutama mitmesõnalisi väljendeid. Liigitusprobleemid tekivad eelkõige siis, kui sõnal puuduvad morfoloogilised ja süntaktilised tunnused. Kui lapse lausung on ilma morfoloogiliste tunnusteta, siis pole ka laiemast kontekstist kasu. (Argus, 2004, 27–29)

Transkriptsioonide käsitsi läbivaatamise puhul hindasin seda, kas tegu on õige lemma, sõnaliigi ja morfoloogiliste kategooriatega. Vale analüüsi saanud sõnade puhul oli väga raske nende sõnaliigilist kuuluvust määrata, sest tihipeale polnud isegi konteksti olemasolul aru saada, millega tegu. Kui see valmistab juba inimesele probleeme, siis pole kahtluski, et analüsaator sellega hakkama saaks, sest tegu on kirjakeelest hälbiva tekstitüübiga. Vale analüüsi saanud sõnadest tegin sagedusloendi ja jaotasin need sõnad 5 erinevasse rühma.

Esimese rühma moodustavad onomatopeetilised sõnad: *viu*, *viuviu*, *nämm*, *amps*, *mämmmämm*, *mämm*, *tapa*, *summ*, *pimm*, *klõps*, *pomm*, *kiiga*, *plaksu-plaksu*, *patsu*, *piiks-piiks-piiks-piiks*, *piiks-piiks*, *pats-pats-pats-pats*, *amps*, *kaak*, *keps*, *sulla*, *kop*, *mõmm*, *põmm*, *kaaga*, *kõps*, *patsu-patsu*, *nämm*, *pisspiss*, *köhi*, *aia*, *tiks*. Nendele sõnadele oli raske sõnaliigilist kuuluvust määrata, mistõttu paigutasin "kirvemeetodil" kõik helijäljenduslikud sõnad ühte rühma.

Teise rühma moodustavad häälitsused ja sõnad, mille tähendusest pole võimalik aru saada ka konteksti olemasolul: *paa*, *eo*, *änn*, *t*, *pupe*, *pigi*, *op*, *muks*, *kookai*, *kaka*, *jää*, *manni*, *öö*, *ämm*, *mm*, *a*, *ä*, *s*, *emm*, *mm*.

Kolmanda rühma moodustavad pärisnimed, mis puuduvad morfoloogilise analüsaatori leksikonist: *Tiibu*, *Triibu*, *Liisu*, *Tups*, *Tuksi*, *Carlos*, *Sirts*, *Sirtsu*, *Annika*, *Antsu*, *Pitsu*, *Alari*.

Neljanda rühma moodustavad sõnad, mis saavad, kas vale lemma või sõnaliigi: *mõmmi*, *siuke*, *venna*, *tudu*, *pai*, *siukse*, *nuku*, *kalli-kalli*, *kalli*, *vot-vot*, *tantsi-tantsi*, *näri-näri*, *mida-mida*, *musi-musi*, *kapp-kapp*, *kapp-kapp-kapp-kapp*, *istu-istu*, *aitab-aitab*, *aluspüksid-aluspüksid-aluspüksid*, *ruttu-ruttu-ruttu*, *väga-väga*, *ja-ja-ja*, *et-et-et*, *tule-tule*, *pisspissi*, *mine-mine*. Siia alla kuuluvad ka sõnad nagu *kuule*, *palun*, *näe*. Need sõnad on siin seetõttu, et need paiknevad verbi ja interjektsiooni piirimail ning oleksid justkui tekkinud täistähenduslike sõnade muutumise teel.

Viienda rühma moodustavad sõnad, mis on oma vormilt vigased (st on läbinud teatud täheteisendused) ja mida CHILDES-i transkriptsioonisüsteemis kodeeritakse [= *explanation*] abil: *ütted* (*ütled*), *ükskold* (*ükskord*), *pilukat* (*pirukas*), *ea* (*hea*), *kah* (*diktofon*), *tee* (*terve*), *kesse* (*kes see*), *emmmee* (*emme*), *auh* (*arvuti*), *te* (*see*), *papa* (*kõndima*), *lau* (*laud*), *kiku* (*diktofon*), *kolla* (*kollane*), *teda* (*seda*), *täh* (*aitäh*), *laua* (*laulma*), *kuku* (*luku*), *kiigu* (*kiik*), *au* (*arvuti*), *olla* (*alla*),

määgu (*mänguasjad*), *kukk* (*trukk* või *raamat*), *koss* (*koos*), *kõrre* (*kõrgel*), *katte* (*kätte*), *kurki* (*kurku*), *noosi* (*joonista*), *kispi* (*küpsis*), *takku* (*traktor*), *ots* (*otsas*), *mammu* (*mari*), *äi* (*ai*), *utu* (*lutt*), *kiika* (*kiikuda*), *kass* (*kastis*), *kapi* (*käbi*), *takka* (*traktor*), *ussi* (*sussid*), *eita* (*ei taha*), *ängi* (*mängib*), *väigi* (*värvi*), *uu* (*õun*), *uksi* (*nutikas*), *toodi* (*joonista*), *tisse* (*televiisori*), *tahta* (*tahan*), *sea* (*see*), *raama* (*raamat*), *puusi* (*pluusi*), *puuniks* (*pruuniks*), *pusti* (*püsti*), *punnu* (*punnis*), *präägust*, *pettu* (*peitu*), *palu* (*palun*), *memme* (*me me*), *märgi* (*värvi*), *mängu* (*mänguasjad*), *mähku* (*mähe*), *laadi* (*lahti*), *kuudi* (*uurima*), *kumme* (*kolm*), *ku* (*diktofon*), *koo* (*koos*), *kõne* (*põnev*), *kombe* (*kombekas*), *kiidu* (*kiisu*), *kii* (*diktofon*), *kat* (*kaks*), *kalju* (*karu*), *kaapi* (*kapi*), *boodi* (*voodi*), *aula* (*laulda*), *auk* (*arvuti*), *aru* (*arvuti*), *ala* (*sajajalgne*), *kipsist* (*küpsist*), *kiisupilt* (*kiisu pilt*), *keti* (*kõdi*), *lutu* (*lutt*), *kumme* (*kummikud*), *sala* (*sajajalgne*), *pillu* (*piilub*), *patt* (*part*), *panni* (*banaanid*), *paa* (*maal*), *olu* (*orav*), *oa* (*orav*), *süü* (*süüa*), *sunni* (*sünnipäev*), *punni* (*punnis*), *käia* (*käima*), *võta* (*võtta*), *õue* (*õues*), *maitse* (*maitsevad*), *koti* (*kott*), *mai* (*mari*), *aua* (*koer*), *voot* (*vot*), *Kate* (*Kattre*).

Tundmatute ning vale analüüsi saanud sõnade hulka saab vähendada analüsaatori allkeelespetsiifilisemaks muutmise teel. Analüsaatori käitumise kohandamiseks tuleks anda sellele sobiv kasutajasõnastik. Kasutajasõnastik on fail, kus igal real on analüüsitav sõnavorm ja selle analüüs. Iga sõna korral kontrollib analüsaator esmalt, kas sõna on kasutajasõnastikus või mitte. Kui on, siis võetakse sealt sõna analüüs, kui ei, siis minnakse morfoloogilist analüüsi tegema. Nii saab kasutajasõnastikku panna sõnu, mida analüsaator muidu analüüsida ei suudaks, ja sõnu, mis peaksid konkreetses tekstis teistsuguse analüüsi saama.

Esimesese rühma ehk onomatopoeetiliste sõnade puhul on raske nende sõnaliigilist kuuluvust määrata, seega tuleks mõelda kasutajasõnastikus uue sõnaliigi defineerimise peale. Teine küsimus on selles, kuidas onomatopoeetilisi sõnu tuvastada? CHILDES-i transkriptsioonisüsteemis on omajagu spetsiifilisi märgendusi, mida on võimalik igale sõnale külge liita. Onomatopoeetiliste sõnade markeerimiseks esitatakse märgendus @o –> *piip*@o, *summ*@o *summ*@o. Sellist tüüpi märgenduse kasutamine teeks onomatopoeetilised sõnad nähtavaks ja oluliselt lihtsustaks nende ekstraheerimist. Kui need sõnad on tuvastatavad, siis on võimalik neid ka automaatselt kasutajasõnastikku lisada. Ainsana on onomatopoeetiliste sõnade markeerimist kasutatud Vija alamkorpuses (781 korral). Seega, praegune olukord näeb ette seda, et onomatopoeetiliste sõnade tuvastamiseks tuleb palju käsitööd teha.

Teise rühma kuuluvad häälitsused ja sõnad, mille tähendusest pole isegi konteksti olemasolul võimalik aru saada. Loomulikult ei tasu juuksekarva lõhki ajada, kuid sellised sõnad on analüsaatori jaoks problemaatilised. Näiteks *t*, *op*, *mm*, *a*, *ä*, *s*, *emm*, *mm* 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 (343x), Zuppingu (3x) ja Kohleri (41x) alamkorpustes. "Sõna" *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 traditsioonilised eesti keele sõnad ja seetõttu saavad omale vale analüüsi.

Kolmanda ja neljanda rühma sõnade valesti analüüsimise vastu ei saa üleskirjutaja midagi teha. Kolmanda rühma sõnad ehk pärisnimed puuduvad analüsaatori kasutajasõnastikust, aga et nende kirjapilt sarnaneb teistele eestikeelsetele sõnadele, siis saavad need ka vale analüüsi. Neljandas 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_*, *mine-mine+0 _V_ Pers Prs Imprt Sg2*. Lisaks on seal sellised sõnad, mis on muidu astmevahelduslikud, kuid on nihutatud astmevahelduseta tüüpi, nagu *mõmmi*, *venna*, *nuku*, mida analüüsitakse genitiivivormideks. Kolmanda ja neljanda rühma sõnade tuvastamiseks ja kasutajasõnastiku täiendamiseks tuleb transkriptsioone palju käsitsi üle vaadata.

Viienda rühma sõnad moodustavad sõnad, mis on oma vormilt vigased (st läbinud teatud täheteisendused) ja mida CHILDES-i transkriptsioonisüsteemis kodeeritakse [= *explanation*] abil. MA KIRJUTAN SIIA KONKREETSE JUTU VEEL JUURDE.

5. Edasine töö

Selles töös jõuti lapsekeele korpuse morfoloogilise analüüsimise esimese katsetuseni. Selles osas annangi ülevaate sellest, kuidas korpusega edasi toimida ehk kuidas oleks võimalik korpuse morfoloogilise analüüsi kvaliteeti tõsta.

Kasutajasõnastik: selleks, et tundmatute ning vale analüüsi saanud sõnade hulka vähendada, tuleks morfoloogilise analüsaatori käitumist muuta. Kasutajasõnastiku tegemiseks oleks tarvis tundmatuid sõnu lähemalt vaadata. Järgnevalt esitan iga alamkorpuse 20 kõige sagedasemat tundmatut sõnavormi (sulgudes olev arv tähistab selle esinemissagedust alamkorpuses).

Kõrgesaar: *vä* (651), *nimodi* (251), *ää* (234), *brmm* (166), *ästi* (129), *nooh* (109), *onju* (104), *präegu* (100), *allo* (97), *Sirlin* (96), *tegelt* (93), *aah* (89), *eksole* (80), *mmh* (69), *mmm* (64), *Demi* (60), *niimodi* (54), *süia* (51), *brumm* (48), *ops* (47).

Vija: *vä* (459), *kessee* (219), *Atsu* (107), *ähäh* (92), *tip* (75), *part_ Toomas* (69), *nooh* (68), *mkmm* (66), *sis* (62), *ää* (54), *ahsoo* (45), *taa* (36), *eksju* (36), *niimodi* (36), *jahh* (35), *tsuhh* (33), *trikstraks* (33), *vat* (32), *mmm* (32), *pss* (30).

Beek: *äta* (758), *vä* (727), *mmh* (339), *mmm* (325), *enna* (287), *Liisbet* (282), *onju* (172), *hõõ* (166), *ääh* (158), *äää* (153), *ää* (120), *aah* (120), *äääh* (110), *ops* (104), *eiä* (86), *õõh* (69), *aaah* (68), *õõ* (65), *mmmm* (65), *ääää* (62).

Kapanen: *Lote* (209), *Martiina* (168), *kaa* (160), *vä* (70), *ää* (68), *mmm* (65), *sis* (46), *äla* (44), *tan* (43), *präegu* (42), *nimodi* (36), *aah* (33), *süia* (33), *puttu* (32), *vata* (31), *ästi* (24), *taa* (24), *ku* (24), *tipa* (21), *näedsa* (20).

Argus: *kaa* (159), *Ninnu* (117), *enna* (41), *drra* (37), *ätaäta* (35), *tiia* (27), *iu* (20), *mäu* (19), *rra* (19), *mämmib* (16), *tiit* (16), *akka* (16), *iiu* (15), *telle* (14), *põrra* (14), *pisi* (13), *atata* (13), *nooh* (13), *iuiu* (12), *plla* (12).

Kohler: *Taimo* (162), *Stella* (143), *Vallu* (71), *Sandor* (70), *allo* (55), *kiss* (54), *aiai* (44), *tak* (39), *onju* (38), *hoppa* (37), *garaazhi* (34), *opa* (31), *Sanna* (29), *miau* (25), *tsuh* (24), *mõmmit* (24), *tit* (22), *tapa* (19), *Kelly* (18), *Sandori* (16).

Zupping: *onju* (83), *äla* (76), *ää* (48), *Eddy* (32), *mõnna* (31), *vä* (29), *mmm* (28), *taan* (27), *jee* (26), *opa* (26), *enna* (25), *daa* (25), *opsti* (24), *pika* (24), *älle* (23), *Ipa* (23), *Krissu* (21), *oih* (20), *numbe* (20), *koppadi* (19).

Iga alamkorpuse sõnavorme vaadates märkame sarnaseid sõnaliike:

- pärisnimed: *Sirlin*, *Atsu*, *Liisbet*, *Lote*, *Martiina*, *Ninnu*, *Taimo*, *Stella*, *Vallu*,

Sandor, *Sanna*, *Kelly*, *Demi*, *Sandori*, *Eddy*, *Ipa*, *Krissu*;

- häälitsused: *ää*, *mmm*, *ätaäta/atata* ja nende erinevad variatsioonid;
- partiklid: *noh*, *aah*, *jahh* ja nende erinevad variatsioonid;
- sõna lühendamine: *vata*, *tegelt*, *vä*;
- kokku liidetud sõnad: *onju*, *eksole*, *ahsoo*, *eksju*, *näedsa*;
- häälduspärane üleskirjutus: *nimodi* ja selle variatsioonid, *präegu*, *enna* (*venna*), *sis* (*siis*), *älle* (*jälle*), *äla* (*ära*), *taan* (*tahan*).

Traditsioonilisest sõnaliigi jaotumisest ei piisa (nt häälitsuste ja onomatopoeetiliste sõnade jaoks), kuid täpsema tegevusplaani jaoks peab tundmatuks jäänud ja ka vale analüüsi saanud sõnu põhjalikumalt analüüsima. Kasutajasõnastiku tegemise puhul tuleb arvestada ka korpuse ja selle sõnavaraga. On tehtud kindlaks, et sõna sagedused järgivad Zipfi seadust. Zipf leidis, et sagedusel ning selle astakul (sõna järjekorranumber sageduste kahanevas reas) sagedussõnastikus on suure dokumendi puhul sõltuvuses ehk siis sõna sagedusel ja selle astaku vahel on funktsionaalne seos, vt (Baayen, 2001, ptk 1). Lihtsamalt lahti seletatuna tähendab see seda, et meil on leksikonis väike hulk sõnu, mis on väga sagedased, ja suur hulk sõnu, mida esineb väga harva. Eelnevalt tõin välja need kõige sagedasemad sõnavormid igas alamkorpuses, aga tabelis 10 on esitatud tundmatuks jäänud sõnavormide (ja neist vaid ühe korra esinevate sõnavormide) koguarv igas alamkorpuses. Tabelist on näha, et vaid ühe korra esinevaid sõnavorme igas korpuses on väga palju, varieerudes 51 ja 74% vahel. Korpuse morfoloogilise analüüsi kvaliteedi parandamisel tuleks sellega ka kindlasti arvestada.

korpus	tundmatuks jäänud sõnavormid	1x tundmatuks jäänud vormid
Vija	1901	1197 63%
Beek	1555	1003 65%
Kapanen	2469	1828 74%
Argus	221	113 51%
Kohler	432	245 57%
Kõrgesaar	3551	2372 67%
Zupping	1355	848 63%

Tabel 10: tundmatute ning 1x esinevate sõnavormide arv ja %

Korpuse märgendamise ja standardiseerimise paremaks muutmine: spontaanse kõne lindistamine ja üleskirjutamine on kahtlemata väga töö- ja ajamahukas , kuid selle töö käigus on kerkinud üles mitmeid üleskirjutamise kitsaskohti (vt ka ptk 4.2).

Üheks suureks kitsaskohaks on sellised sõnad, mis on läbinud teatud täheteisendused. Osad transkribeerijad on lapse poolt öeldut kirjakeelsemaks muutnud, osad jällegi kasutavad üleskirjutamisel palju kuuldeortograafiat (ja sedagi ebajärjepidevalt). Hennoste kirjutab, et suulise kõne transkriptsiooni koostamisel on kaks printsiipi. Esimene printsiip 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 praktilised. See tähendab seda, et märgendada tuleb neid nähtuseid, mida uurijal on tarvis, kuid et see üldist pilti üle ei küllastaks. (Hennoste, 2002, 92–93) On arusaadav, et iga transkriptsioon on üles kirjutatud vastavalt uurija eesmärkidele, kuid siinkohal tuleks mõelda sellele, kuidas selline üleskirjutamise viis mõjutab sellele järgnevat analüüsi ja töötlust. Morfoloogilise analüüsi seisukohalt on hetkeolukord selline, et suur osa vale analüüsi saanud (vt alaptk 4.2 esitatud viiendat sõnarühma) ja tundmatuks jäänud sõnu sõltuvad paraku üleskirjutamise viisist. Küsimus ei ole selles, et korpuse tegija peab kvaliteetsema analüüsi nimel loobuma autentsusest ja praktilisusest, vaid selles, kas transkribeerija on nõus natuke rohkem vaeva nägema ja valmis üles märkima nii vigaseid (ehk mida laps

tegelikult ütles) kui õigeid (ehk mida öelda taheti, kirjakeelne variant) vorme.

Teine aspekt (vt ptk 4.2) on onomatopoeetilised sõnad ja häälitsused. CHILDES-i konventsioonide järgi on iga sõna külge võimalik liita spetsiaalseid märgendusi: @o on onomatopoeetilise sõna markeerimiseks, @x saab kasutada sõna välja jätmiseks, @k abil saab silpe markeerida, @l abil saab ühte häälikut markeerida, @c ehk lapse väljamõeldud vormi markeerimiseks, @z abil saab transkribeerija kasutada oma defineeritud märgendust jpm. Hetkeolukord näeb ette, et kasutajasõnastikku täiendatakse käsitsi, kuid spetsiaalsete märgenduste kasutamine võimaldaks seda automaatselt teha.

Korpuse ühestamine: korpus on praegusel kujul ühestamata, see tähendab, et kõikvõimalikud analüüsid on alles jäetud. Pikemas perspektiivis on kindlasti plaan teha nii, et iga sõna saaks vaid ühe analüüsivariandi, kuid see nõuab palju käsitsi üle vaatamist.

Kokkuvõte

TEEN TÄNA, OLENEB MUIDUGI ANALÜÜSI PEATÜKKIDEST.

Summary

Lisad

Lisa 1. Vija

Lapse nimi	Vanus	Sugu	Sessioonid	Lapse sõnad	Hoidja sõnad	KOKKU
Andreas	1;7-1;11	p	7	2845	8521	11366
	2;0-2;8		37	41498	59272	100770
	3;0-3;1		30	66038	48137	114175
KOKKU			74	110381	115930	226311

Lisa 2. Argus

Lapse nimi	Vanus	Sugu	Sessioonid	Lapse sõnad	Hoidja sõnad	KOKKU
Hendrik	1;8-1;11	p	5	566	1963	2529
	2;0-2;5		12	3654	7190	10844
KOKKU			17	4220	9153	13373

Lisa 3. Beek

Lapse nimi	Vanus	Sugu	Sessioonid	Lapse sõnad	Hoidja sõnad	KOKKU
Liisbet	0;9-0;11	t	6	1450	11487	12937
	1;0-1;2		5	1143	10463	11606
	2;0-2;5		9	7571	27022	34593
KOKKU			20	10164	48972	59136

Lisa 4. Kapanen

Lapse nimi	Vanus	Sugu	Sessioonid	Lapse sõnad	Hoidja sõnad	KOKKU
Martina	1;3-1;11	t	6	7302	17791	25093
	2;1-2;7		4	6831	9831	16662
	3;1		1	1805	2115	3920
KOKKU			11	15938	29737	45675

Lisa 5. Kõrgesaar

Lapse nimi	Vanus	Sugu	Sessioonid	Lapse sõnad	Hoidja sõnad	KOKKU
Andri	11;7-11;9	p	2	4253	4445	8698
Arabella	11;8	t	1	200	2439	2639
Artur	1;4	p	1	94	3540	3634
Gregory	6;6	p	1	3435	4570	8005
	7;1-7;8		2	4501	7559	12060
	8;4-8;10		3	4840	5823	10663
	9;7-9;8		2	3927	4125	8052
	10;5		2	4822	4740	9562
Harley	4;0-4;1	p	5	1357	3269	4626
	7;2		3	3589	3208	6797
	10;1-10;2		2	3232	3355	6587
	11;0-11;11		4	7838	6716	14554
	12;5		1	2003	1344	3347
	13;2-13;3		2	4290	4261	8551
	14;0-14;1		2	4976	4665	9641
	4;0	t	2	412	870	1282
Hellyn	8;7	t	1	1438	2066	3504
Jaana	2;5	t	1	1447	1841	3288
Kaisa	5;8-5;9	t	2	3256	4690	7946
Mia	2;3	t	1	1643	5368	7011
Olivia	3;2	t	1	1275	2882	4157
Ruuben	1;3-1;4	p	2	777	3544	4321
	2;2		1	936	3355	4291
	3;6		1	1259	2669	3928
Sirlin	1;3	t	1	19	2348	2367
KOKKU			46	65819	93692	159511

Lisa 6. Zupping

Lapse nimi	Vanus	Sugu	Sessioonid	Lapse sõnad	Hoidja sõnad	KOKKU
Linda	1;3-1;11	t	9	3677	15191	18868
	2;0-2;11		12	6785	13114	19899
	3;0		1	542	1052	1594
	4;2		1	628	1472	2100
KOKKU			23	11632	30829	42461

Lisa 7. Kohler

Lapse nimi	Vanus	Sugu	Sessioonid	Lapse sõnad	Hoidja sõnad	KOKKU
Anna	1;10-1;11	t	4	550	4298	4848
	2;0-2;1		3	645	3454	4099
Carlos	1;7-1;10	p	9	1797	6809	8606
Helen	1;1-1;10	t	7	551	7745	8296
Henri	2;2-2;3	p	3	633	2612	3245
Mari	2;5-2;8	t	7	2455	7850	10305
Sandor	1;2-1;10	p	7	1219	9445	10664
	2;2	p	3	1374	4564	5938
Stella	0;11	t	1	6	295	301
	1;0-1;6		8	287	6629	6916
Taimo	1;5-1;11	p	9	536	6958	7494
KOKKU			61	10053	60659	70712

Kasutatud kirjandus

Aguado-Orea, J. & Pine, J. M. (2015). Comparing different models of the development of verb inflection in early child spanish, *PLoS ONE* **10**(3): 1–21.

Argus, R. (2004). Imitatiivide kohast lastekeeles: reduplikatsioonist, morfoloogiast ja sõnaliigilisest ambivalentsusest, *Eesti Rakenduslingvistika Ühingu aastaraamat* **1**: 19–34.

Argus, R. (2007). Eesti lastekeelekorpuse morfoloogilisest märgendamisest, *Tallinna ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine* pp. 65–86.

Argus, R. (2008a). *Eesti keele muutemorfoloogia omandamine.*, doktoritöö, Tallinna Ülikool.

Argus, R. (2008b). Kuidas eesti laps vormimoodustuse omandab., *Oma Keel* (16): 17–26.

Argus, R. & Kõrgesaar, H. (2014). Sõnaliigid eesti lapse kõnes ja lapsele suunatud kõnes., *Eesti Rakenduslingvistika Ühingu aastaraamat* **10**: 37–53.

Baayen, R. H. (2001). *Word Frequency Distributions.*, Kluwer Academic Publishers.

Behrens, H. (2008). Corpora in language acquisition research: History, methods, perspectives, *Corpora in Language Acquisition Research: History, methods, perspectives* pp. 11–30.

Burnard, L. (2014). *What is the Text Encoding Initiative? How to add intelligent markup to digital resources*, OpenEdition Press.

Chatter tarkvara (2016). `http://talkbank.org/software/chatter.html`. 18.04.2016.

CHILDES (2016). Childes-i andmebaas, `http://childes.psy.cmu.edu/data/`. 06.05.2016.

Choi, S. & Gopnik, A. (1995). Early acquisition of verbs in korean: a cross-linguistic study., *Journal of Child Language* **22**: 497–529.

EKK (2007). *Erelt, Mati and Erelt, Tiiu and Ross, Kristiina. Eesti keele käsiraamat*, Eesti Keele Sihtasutus.

etTenTen (2015). ettenten, `http://www2.keeleveeb.ee/dict/corpus/ettenten/about.html`. 05.03.2016.

Gentner, D. (1982). Why nouns are learned before verbs: linguistic relativity vs. natural partitioning., *Language development Vol 2: Language, thought and culture* pp. 301–334.

Gillis, S. (2014). *Child Language Data Exchange System*, pp. 74–78.

Hennoste, T. (2002). Suulise kõne uurimine ja sõnaliigi probleemid, *Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised* **4**: 56–73.

Kaalep, H.-J., Muischnek, K., Müürisep, K., Rääbis, A. & Habicht, K. (2000). Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? eesti keele testkorpuse morfosüntaktilise märgendamise kogemusest., *Keel ja Kirjandus* **9**: 623–633.

Kaalep, H.-J. & Vaino, T. (2000). Teksti täielik morfoloogiline analüüs lingvisti töövahendite komplektis, *Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1* pp. 87–101.

Korpused ja keelekogud (2015). `http://www.keel.ut.ee/et/keelekogud`. 05.03.2016.

Krajewski, G., Theakston, A. L. & Lieven, E. V. M. (2012). Productivity of a polish child's inflectional noun morphology: a naturalistic study, *Morphology* pp. 9–34.

Kõrgesaar, H. (2009). Hoidjakeelele omastest joontest., *Oma Keel* (2): 28–37.

Kõrgesaar, H. & Kapanen, A. (2015). Kui lapsega ei räägi üksnes ema: valik termineid eesti laste- ja hoidjakeele kohta., *Eesti Rakenduslingvistika Ühingu aastaraamat* **11**: 177–188.

Laing, C. E. (2014a). A phonological analysis of onomatopoeia in early word production., *First language 34* pp. 387–405.

Laing, C. 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* .

Leech, G. (2005). Adding linguistic annotation, `http://www.ahds.ac.uk/creating/guides/linguistic-corpora/chapter2.htm`. 11.05.2016.

Lieven, E. V. M. (2010). Input and first language acquisition: Evaluating the role of frequency, *Lingua 120* pp. 2546–2556.

MacWhinney, B. (2016). Part 1: The chat transcription format. the childes project: Tools for analyzing talk – electronic edition, `http://childes.psy.cmu.edu/manuals/CHAT.pdf`. 18.04.2016.

MacWhinney, B. & Snow, C. (1985). The child language data exchange system., *Journal of Child Language 12* pp. 271–296.

McEnery, T. & Hardie, A. (2011). *Corpus Linguistics. Method, Theory and Practice*, Cambridge University Press.

Muischnek, K. (2015). Keelekorpused – sama mitmekesised kui keel ise, *Oma Keel* **1**: 37–44.

Muischnek, K., Kaalep, H.-J. & Sirel, R. (2016). Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile, `http://www.keeleveeb.ee/dict/corpus/comments/uue_meedia_morf.pdf`. 11.05.2016.

Muischnek, K., Orav, H., Kaalep, H. & Õim, H. (2003). Eesti keele tehnoloogilised ressursid ja vahendid. arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara, pp. 1–86.

Muischnek, K. & Vider, K. (2004). Sõnaliigituse kitsaskohad eesti keele arvutianalüüsis, *Eesti Rakenduslingvistika Ühingu aastaraamat* **1**: 99–114.

Newport, E. L., Gleitman, H. & Gleitman, L. R. (1977). "mother, i'd rather do it myself: some effects and noneffects of maternal speech style", *Talking to children.* pp. 109–149.

Ochs, E. (1979). Transcription as theory, *Developmental pragmatics* pp. 43–72.

Orusalu, S. (2008). *Lastega suhtlemise erisõnavara.*, diplomitöö, Tartu Ülikool.

Seene, K. (2015). *Verbi 'minema' omandamine: semantiline ja süntaktiline analüüs.*, bakalaureusetöö, Tartu Ülikool.

Soderstrom, M. (2007). Beyond babytalk: Re-evaluating the nature and content of speech input to preverbal infants, *Developmental Review 27* pp. 501–532.

Tardif, T. (1996). Nouns are not always learned before verbs: evidence from mandarin speakers' early vocabularies., *Journal of Child Language* **32**: 492–504.

Theakston, A. L., Lieven, E. V. M., Pine, J. M. & Rowland, C. F. (2002). Going, going, gone: the acquisition of the verb 'go'., *Journal of Child Language 29* pp. 783–811.

Tomasello, M. & Stahl, D. (2004). Sampling children's spontaneous speech: how much is enough?, *Journal of Child Language 31* pp. 101–121.

Vabamorfi morfoloogia-leksikon (2016). `https://raw.githubusercontent.com/Filosoft/vabamorf/master/doc/morfi_leksikoni_kirjeldus.html/`. 16.05.2016.

Vaik, K. (2014). *Eesti keele muutemorfoloogia omandamisest 'wug' katse põhjal.*, bakalaureusetöö, Tartu Ülikool.

Vider, K. (1995). *2–3-aastaste eesti laste sõnavara.*, diplomitöö, Tartu Ülikool.

Vihman, M. M. & Vija, M. (2006). *The acquisition of verbal inflection in Estonian.*

Vihman, V.-A. (2015). Pick it up: a look at referential devices in estonian child-directed speech, *Eesti ja soome-ugri keeleteaduse ajakiri 6–2* pp. 63–85.

Wexler, K. & Culicover, P. W. (1980). *Formal principles of language acquisition.*, MIT Press.

XML Tutorial (2016). `http://www.w3schools.com/xml/default.asp`. 18.04.2016.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks.

Mina, Kristiina Vaik (sünnikuupäev: 16.12.1990)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Morfoloogiliselt märgendatud eesti lapsekeele korpus", mille juhendajad on Heiki-Jaan Kaalep ja Virve-Anneli Vihman,
   (a) reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace'i lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   (b) 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.

Tartu, 31. mai 2016. a.