

TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Kristiina Vaik

Eesti morfoloogiliselt märgendatud lapsekeele korpus
Magistritöö

Juhendajad: Heiki-Jaan Kaalep ja
Virve-Anneli Vihman

TARTU 2016

Sisukord

Sissejuhatus	3
1. Lapsekeel ja sellele iseloomulikud jooned	4
1.1. MINGI PEALKIRI	4
2. Suuline keel ja sellele iseloomulikud jooned	9
3. Korpused	10
3.1. Mis on korpus?	10
3.2. Eesti keele korpused	11
3.3. Korpuse märgendamine	12
3.4. Mis on XML?	13
3.5. Sõnaliikide kitsaskohad	15
4. CHILDES ja eesti keele alamkorpused	17
4.1. CHILDES	17
4.2. Mida teistes keeltes tehtud vms...	19
4.3. Alamkorpuste standardiseerimise probleemid	19
4.4. Eesti keele alamkorpused	22
4.5. Tabelid	26
4.5.1. Kõrgesaar	26
4.5.2. Kapanen	29
4.5.3. Beek	31
4.5.4. Kohler	33

4.5.5. Vija	36
4.5.6. Argus	38
4.5.7. Zupping	40
4.5.8. Kõik korpused	42
5. Morfoloogiliselt märgendatud lapsekeele korpus	43
5.1. Tööprotsess	43
5.1.1. <i>Talkbanki</i> skeema	45
5.2. Morfoloogilise info lisamise protsess	47
6. Kokkuvõte	48
Kasutatud kirjandus	50

Sissejuhatus

SEE KOKKUVÕTE OLI SEMINARITÖÖ JAOKS. SEEKA, SEE POLE SEE “PÄRIS” KOKKUVÕTE. JA PEALKIRJAD ON KA ALLES ÜSNA ALGELISED.

Seminaritöö eesmärgiks on teha lühike sissejuhatus magistratöö vaid ühe praktilise väljundi kohta, milleks on eesti morfoloogiliselt märgendatud lapsekeelete korpus. Korpuse tekstdid pärinevad CHILDES-i eesti lapsekeelete alamkorpustest. Morfoloogiliselt märgendatud lapsekeelete korpuse loomine oleks väga vajalik, sest CHILDES-i tööriistad ei võimalda teha eesti keele jaoks automaatset morfoloogilist analüüsni.

Töö koosneb kahest peatükist. Esimeses peatükis räägitakse lähemalt korpusest, selle liigitusvoimalustest ja arengutendentsidest. Lisaks antakse lühiülevaade sellest, mida korpuse märgendamisel silmas tuleb pidada, ja millised korpused juba Tartu Ülikoolis olemas on. Teises peatükis kirjutatakse morfoloogiliselt märgendatud lapsekeelete korpuse loomise probleemidest. Tutvustatakse CHILDES-i korput ja kuidas on seal esindatud eesti keele alamkorpused. Lõpuks tehakse väike ülevaade alamkorpuste standardiseerimise probleemidest.

1. Lapsekeel ja sellele iseloomulikud jooned

Siin: Reili Argus ja võimalusel ka muudki.

1.1. MINGI PEALKIRI

Eesti keele morfoloogia on rikkalik, kuid selle omandamine on keerukas protsess. Keele omandamisel tuleb lisaks reeglipärastele mallidele omandada ka ebareeglipärased mallid ning teada, kuidas toimuvad tüvesisesed muutused (astme- ja lõpuvaheldus) ning milliseid morfoloogilisi formative (tunnused ja lõpuid) tuleb tüve külge lisada. Konstruktivistlik lähenemine jagab morfoloogiasüsteemi omandamise kolme perioodi (Dressler 2003: 9-10) VIIDE:

- Premorfoloogiline omandamine- etapp, millal lapse keeles esinevad morfoloogilised muuted, kuid morfoloogiline süsteem pole lahus kognitiivsetest struktuuridest;
- Protomorfoloogiline periood- etapp, mil hakkab arenema morfoloogiline produktiivsus. Laps loob omale grammatika, mis on täiskasvanutest reeglipärasm ja pisem, kuid siiski abistab omandamist siis, kui muutemorfoloogia süsteem pole veel selge. Toimub esimeste reeglite ja analoogiaseoste loomine. Alguses keskendutakse erinevatele vormidele, mis muutuvad teatud kategooriate prototüüpideks;
- Täiskasvanute morfoloogia omandamine- etapp, mil keelesüsteemi allmoodulid hakkavad omavahel koostööd tegema. Sel ajal hakkavad lapsel aktiivselt välja kujunema muutmissüsteemid, sealhulgas liitmine ja tuletamine.

Esimesed poolteist eluaastat tegelevad lapsed sisendkeele desifreerimisega ehk püütakse öeldule tähendust omistada. Morfeemide hulk keeles on väga suur, seega pole laps võimeline korraga kõiki morfeeme ära õppima ning selle töttu keskendutakse pigem sagedasematele ja esilduvamatele morfeemidele. Olulised tegurid morfeemide valiku juures on sagedus, segmenteeritavus, morfeemi asend tüves ning selle allomorfide esinemise seaduspärasused. (Argus 2004: 16, Argus 2008: 19) VIIDE Esmalt peab laps selgeks saama kust algab ja lõpeb sõna. Esmalt omandatakse lühikesi lahtiseid silpe, kuna neid on lihtsam häädada ja segmenteerida kui kinniseid. Kõnejada segmenteerimist ja silbipiiri ära tundmist hõlbustab reduplikatsioon, mis on hoidjakeelele väga omane (ta-da, ai-ai jne). Hoidjakeele rutiinsed väljendid aitavad lapsel aru saada, kust algab samasugune foneetiline realisatsioon. Reduplikatsioon aitab mõista sõnu kui mitmest osast koosnevast tervikust, mis

omakorda aitab hilisemas etapis tajuda sõnavormis esinevat muutumatut (tüvi) ja muutuvat (tunnused ja lõpuud) osa. Reduplikatiivsed sõnad hõlbustavad morfoloogia omadamist, sest tunnuseta vorme hakatakse tasapisi asendama morfoloogilisi elemente sisaldavate üksustega. Reduplikatsioon aitab omandada ka eesti keelele iseloomulikku kõnetakti ehk trohheust, mille puhul ühest lahtisest silbist tekita takse kahesilbilise keeleüksus, kus vahelduvad rõhuline ja rõhuta silp. Enne kahe aastaseks saamist on lapse kõnes enamik keeleüksuseid trohheilise rütmiga. (Argus 2008a: 19-20) VIIDE

Premorfoloogilisel etapil omandavad lapsed vorme terviklike üksustena, ilma et toimuks tüve ja tunnuse analüüsist. Sel perioodil esineb laste kõnes kõige rohkem tunnusteta kahesilbili sõnu. Pikemates vormides, kas loobutakse rõhutust ehk viimasesilbist (nt *sitikas* > *siti*, *traktor* > *takku/takka*), *rebane* > *repa* või jäetakse käändelõpp ära ja kasutatakse selle asemel genitiivi tüve (nt *poti* > *potile*). Kahesilbliste vormide kasutust võib vaadata nii eelistuse kui ka piiranguna. Piirang selles mõttes, et laps ei pruugi olla suuteline hääldamata pikemates sõnades teatud rõhutut silpi. Eelistus selles mõttes, et laps eelistab kahesilbili sõnu, kuna need on sisendkeeles kõige sagedasemad. Nii eesti kui ka soome keeles on lapsekeelete uurijad märganud, et on perioode, mil laps pikendab ühesilbilise sõna kahesilbili seks, nt *lutt* > *luti*, *pai* > *paia*. Näiteid on ka sellest, et liitsõnade omandamisel püütakse eelistada kahesilbili sõnu, nt *jõuluvana* > *jõvvu*. Vahetult enne protomorfoloogilist etappi lõpeb kahesilbliste sõnade periood. (Argus 2008a: 20–21) VIIDE

Kinniste silpide omandamine toimub nende hääldamise raskuse tõttu hiljem. Eesti keeles toimub kinniste esimeste silpide omandamine kinnistest järgsilpidest varem. Leidub näiteid, kus kinnise järgsilbiga sõnu mugandatakse trohheiliseks, nt *viiul* > *villu/illu*, *ämber* > *ämpu*, *põrr* > *põrra*. Premorfoloogilisele etapile on iseloomulik, et lapse kõnes kinnise järgsilbi lõppkonsonant on ära jäetud, kuna laps ei kasuta selliseid käände- ja pöördelõppe, mis muudaksid järgsilbi kinniseks. Näiteks, partiidiivi lõpp -t jäääb hääldamata (*küpsist* > *kispi*), mitmuse nominatiivi lõpp -d jäääb hääldamata (*mammud* > *mammu*), inessiivi lõpp -s jäääb hääldamata (*vannis* > *vanni*, *aias* > *aia*). Konsonandi hääldamine sõltub paljuski konsonandi positsioonist: sõnaalguline konsonant jäääb tavaliselt alles, kuid sõnalõpuline konsonant mitte. Kinnise järgsilbi hilisem omandamine on ootuspärane, kuna eesti keeles on kahesilbilised lahtised lõppsilbiga sõnad väga sagedased. (Argus 2008a: 22) VIIDE

Reili Argus (2008a: 23) VIIDE 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 premorfoloogilisel perioodil,

mil ei ole tunnused ega lõpuud veel omandatud. Vältevaheldus omandatakse varakult, kuna vältevahelduslikud sõnad on sisendkeeles sagedes ning vältetele opositsioonidel on grammatiske tähinduste eristamisel tähtis roll. Näiteks nõrga- või tugevaastmelised sõnad eristavad lapse jaoks grammatiske tähindusi- valdaja ja objekt või objekt ja asukoht. (Argus 2008a: 23) VIIDE

Lõpuvaheldusega seotud vigu esineb üleminekul premorfoloogiliselt perioodilt protomorfoloogilisele perioodile. Enim valmistab probleeme II-välteliste konsonant-lõpuliste sõnade vormimoodustus. Peamiselt just selliste fonoloogiliselt keeruliste sõnadega (*el-*, *er-*lõpulised), kus laps väldib kolmest konsonandist koosnevat kaashäälikuühendit (nt **numbert*, **numberit* 'numbrit'; **kahvelga*, **kahveliga* 'kahvliga'). Raskusi on ka *s*-lõpuliste sõnade vormimoodustusega (nt **kärbese* 'kärbse'; **võõraseid* 'võõraid'). Vigu esineb ka *ne*-liiteliste sõnade vormimoodustusega, näiteks nominatiivkujulise tüve *rebane* asemel kasutatakse **reba* ja genitiivivormina **rebali*. Olgugi, et Arguse poolt vaatluse all olevad lapsed olid vältevahelduse varakult omandanud, siis laadivaheldus koos lõpuvaheldusega nagu *V > me* põhjustas mõningaid raskusi. Näiteks oli laps ära õppinud vormi *juhtmed* ning järgmisste vormide käänamine toimubki selle ühe vormi analoogial (ehk **juhtme* 'juhe'; **juhtmet* 'juhet') ning hilisemas materjalis on näha, et lapse keelekasutusse tekib vorm **juhte* 'juhtme'. See näitab, et laps on tajunud, et *m*-häälik kaob, kuid ei oska seda veel õigesti kasutada. (Argus 2008a: 23–24) VIIDE

Lõpuvaheldus pole nii süsteemne kui astmevaheldus ja selle omandamine on raske, kuna tüvevahelduslike vormide puhul tuleb lõpfoneemide järjestust vahetada. Lihtsaim viis sõnade moodustamiseks on lõpuhäälikute lisamine, kuid vahel üldistatakse muutesufiksit ka sõnadele, kus see ei ole normikohane, nt *kauss*: **kausit* (partitiivi läbipaistva lõpu *-t* üldistamine), *tühi*: **tühja-sse* (illatiivi läbipaistva lõpu *-sse* üldistamine). Eesti keeles on noomenitel lõpuvaheldusmalli rohkem kui pöördsõnadel, mistõttu ei valmista pöördsõnade lõpuvaheldusmallid lastele ka probleeme. Näiteks sellised tüvevaheldused nagu *sööme*: *süüa*, *lööb*: *lüüa*, *ei pea*: *pidime* on omandatud veatult. (Argus 2008a: 24, 26–27, 2008b: 20) VIIDE

Reduplikatiivsuse kõrval on morfoloogia süsteemi omandamisel tähtis roll ka deminutiivtuletustel, mille käigus nihutatakse astmevaheldusega sõnad astmevaheldusesta muuttüipi. Näiteks *kiisu* ja *kutsu* saavad selle nihke töttu partitiivi läbipaistva lõpu *-t*: *kiisu-t*, *kutsu-t*. Siinkohal on hõlbustav tegur ka vormihomonüümia, sest tuletatud sõnad nihkuvad sellisesse muuttüipi, kus nominatiivi- ja genitiivivormid on homonüümsed. Väidetakse, et samakujulised vormid soodustavad muuteparadigmade omandamist, kuid see võib olla ka pidurdavaks teguriks. Nimelt võib laps hakata käänama ka *pesa*-tüüpi sõnu astmevahelduseta muuttüübi järgi, nt **mu-na-t*, **saba-t*. (Argus 2008a: 25, 2008b: 19–20) VIIDE

Lähemalt võiks natuke rääkida ka muutemorfoloogia arengust sõnaliigiti. Argus kirjutab (2008a), et morfoloogiasüsteemi omandamine toimub sõnaliigiti erineva kiirusega ja et mõnes keeles omandatakse noomeni morfoloogia 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. Verbid on semantiliselt keerukamad ja on rohkem seotud keele süntaktilise struktuuriga. (Argus 2008a: 37) VIIDE Kasutuspõhise lähenemise järgi kontrueerivad lapsed grammatika sellest, mida nad kuulevad. Näiteks kui keeles on ülekaalus substantiivid, siis ainult sellepärast, et substantiivide esinemissagedus on sisendkeeles suur.

«««< Updated upstream

Melissa Bowerman annab oma loengusarjas “Ten lectures on language, cognition and language acquisition” (2010 VIIDE) ülevaate sellest, kuidas lapsed õpivad kaardistama keelelisi vorme (sõnad, morfeemid, sõnajärg konstruktsioonid jpm) ja nende tähendusi. Sellele küsimusele läheneti keeltevahelise võrdluse perspektiivist. Välja tuli see, et mõned keeled on verbi-“sõbralikumad” kui teised. Näiteks inglise hoidjakeeles domineerivad noomenid, kuid on keeli, kus noomenite ja verbide kasutus on ühtlasemalt jaotunud, nt korea ja mandariini keel ning tseltali ja tsotsili keel. Nendes keeltes on objekti informatsioon peitetud verbidesse. Näiteks tseltali ja tsotsili keeles on mitu verbi söömise tähistamiseks ja vastava verbi valik sõltub sellest, mida süüakse. Seega verbid kannavad endas objekti kohta palju informatsiooni ja seetõttu muutuvad noomenid diskursuses vähem oluliseks. See, millise sõnaliigi muutemorfoloogia omandamine kiiremini toimub, oleneb sisendkeele eripärast ja on vägagi keelespetsiifiline. (Bowerman) VIIDE. ===== JAMA: Melissa Bowerman annab oma loengusarjas “Ten lectures on language, cognition and language acquisition” (2010 VIIDE) ülevaate sellest, kuidas lapsed õpivad kaardistama keelelisi vorme (sõnad, morfeemid, sõnajärg konstruktsioonid jpm) ja nende tähendusi. Sellele küsimusele läheneti keeltevahelise võrdluse perspektiivist. Välja tuli see, et mõned keeled on verbi-“sõbralikumad” kui teised. Näiteks inglise hoidjakeeles domineerivad noomenid, kuid on keeli, kus noomenite ja verbide kasutus on ühtlasemalt jaotunud, nt korea ja mandariini keel ning tseltali ja tsotsili keel. Nendes keeltes on objekti informatsioon peitetud verbidesse. Näiteks tseltali ja tsotsili keeles on mitu verbi söömise tähistamiseks ja vastava verbi valik sõltub sellest, mida süüakse. Seega verbid kannavad endas objekti kohta palju informatsiooni ja seetõttu muutuvad noomenid diskursuses vähem oluliseks. See, millise sõnaliigi muutemorfoloogia omandamine kiiremini toimub, oleneb sisendkeele eripärast ja on vägagi keelespetsiifiline. (Bowerman) VIIDE. »»»> Stashed changes

Argus ja Kõrgesaar uurisid sõnaliikide jaotumist ja esinemissagedust lapse ja täiskasvanu vahelistes spontaansetes igapäevavestlustes. Sõnaliikide jaotusena kasutati traditsionilist liigitust, onomatopoeetilisi sõnu ning ebaselgeid sõnu ja üneeme arvestati eraldi.

KIRJUTA: Sõnaliikide jaotumisest

?

Hodjakeele erijooned

Hodjakeel ja lapsekeel kui sotsiolektid

2. Suuline keel ja sellele iseloomulikud jooned

Peamiselt Tiit Hennostelt.

3. 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.

3.1. Mis on korpus?

Enne arvutite kasutuselevõttu mõeldi keeleteaduses keelekorpuse all kui 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 keest ja iseloomustaksid keele hetkeseisu või muutumist. Tekste valitakse korpusesse teatud kriteeriumite alusel ja need esinevad standardses elektroonilises formaadis. (Muischnek et al., 2003, 9)

Korpuste põlvkonnad:

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. (Muischnek, 2015b)

Esimesed elektroonilised tekstikorpused olid Browni ja Lancaster-Oslo/Bergeni (LOB) korpused. Nendesse korpusestesse lisatud tekstdid 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: *Bri-*

tish 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)

3.2. 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, tekstdid 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 tekstiklassi. 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. Tekstiklassid ei ole kindlaks määratud. 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, 2015a, 38) Eesti keele kolmanda põlvkonna korpuseks on *eTen-Ten*, mis koosneb on 270 miljonist sõnast 686000 veebilehelt. (*eTenTen*, 2015)

Paralleelkorpus on korpus, mis sisaldab sama teksti vähemalt kahes keeles, mille üksused on paralleelistatud. *Parallelistamine* tähendab paralleelteksti üksteisse tõlkeks olevate osade märgendamist ehk pannakse paika, et millised laused, osalaused, fraasid on omavahel vastavuses. Tuntuim paralleelkorpus maailmas on Kanada *Hansard*, mis koosneb inglise- ja prantsusekeelsetest parlamentidebattidest. Tartu Ülikoolis on tehtud eesti-inglise paralleelkorpus. Eesti keelt sisaldavaid paralleelkorpuseid on vähe. Eesti keelt sisaldab Soome paralleelkorpus *SCLOMB*, mis sisaldab Läänemere-äärsete keelte tekste ja nende tõlkeit teistesse Läänemere-äärsetesse keeltesse. Keeletehnoloogia seisukohalt on paralleelkorpus väga oluline keeleressurss, nt saab paralleelkorpsi kasutada masintõlkesüsteemide loomisel ja sõnastike automaatsel genereerimisel. (Muischnek et al., 2003, 17–18)

KAS MA PEAKS NEIST KORPUSTEST PIKEMALT KIRJUTAMA??

Eesti keele erikorpused:

1. suulise kõne korpus;
2. spontaanse kõne foneetiline korpus

3. dialoogikorpus;
4. murrete korpus;
5. vana kirjakeele korpus;
6. foneetikakorpus jne. (*Korpused ja keelekogud*, 2015)

3.3. Korpuse märgendamine

SELLES ALAPEATÜKIS MA KAHTLEN, SEST SEE ON JUSTKUI LIIGA COPY-PASTE. JA MUL POLE TEGELIKULT JU TARVIS SÜNTAKTILISEST JA SEMANTILISEST TASANDIST RÄÄKIDA, FOOKUS ON SIISKI MORFOLOOGILISEL ANALÜÜSIL. PIGEM TEEKS ALAPEATÜKI MORFANALÜSAATORIST JA SELLE TEHNILISEST KÜLJEST.

Korpustest on kasu siis, kui vajalik info on sealt lihtsalt kättesaadav. Selleks, et korpus ei jäeks lihtsalt elektrooniliste tekstide arhiiviks, oleks tarvis korpusesse esmalt interpretatiivset infot lisada. Seda protsessi nimetatakse *korpuse märgendamiseks*. Korpuse märgendamise etapis lisatakse tekstile infot nende ülesehituse kohta ning morfoloogilise, süntaktilise, semantilise jne analüüsi tulemused. Korpuse märgendamise juures on vajalik selgeks teha, mida (sisu) ja kuidas (vorm) märgendada. Märgendamist saab teha automaatselt, käsitsi või neid mõlemaid kombinererides (poolautomaatselt). Märgendamist alustatakse esmalt tehniline 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). Enim levinud tasemed on morfoloogiline ja süntaktiline märgendamine. (Muischnek et al., 2003, 12–13)

Morfoloogiline märgendamine annab iga sõna jaoks infot selle lemma ehk algvormi, sõnaliigi ja morfoloogiliste kategooriate kohta (kääändsõnal arv ja käane, tegusõnal pööre, tegumood, aeg, kõneviis, kõneliik). Morfoloogilist märgendamist saab teha automaatselt (morfoloogiline analüsaator ja ühestaja) või poolautomaatselt. Seejärel toimub morfoloogiline ühestamine. *Morfoloogiliseks ühestamiseks* nimetatakse protsessi, kus sõnale on morfoloogilise analüüsi käigus määratud kõik võimalikud interpretatsioonid ja ühestaja eesmärk on nendest interpretatsioonidest välja valida antud konteksti sobiv analüüs. (Kaalep & Vaino, 2000)

Süntaktiline märgendamine põhineb kitsenduste grammatika formalismil, mille käigus lisatakse alguses igale sõnale kõik analüüsvariandid ja hiljem eemalda-

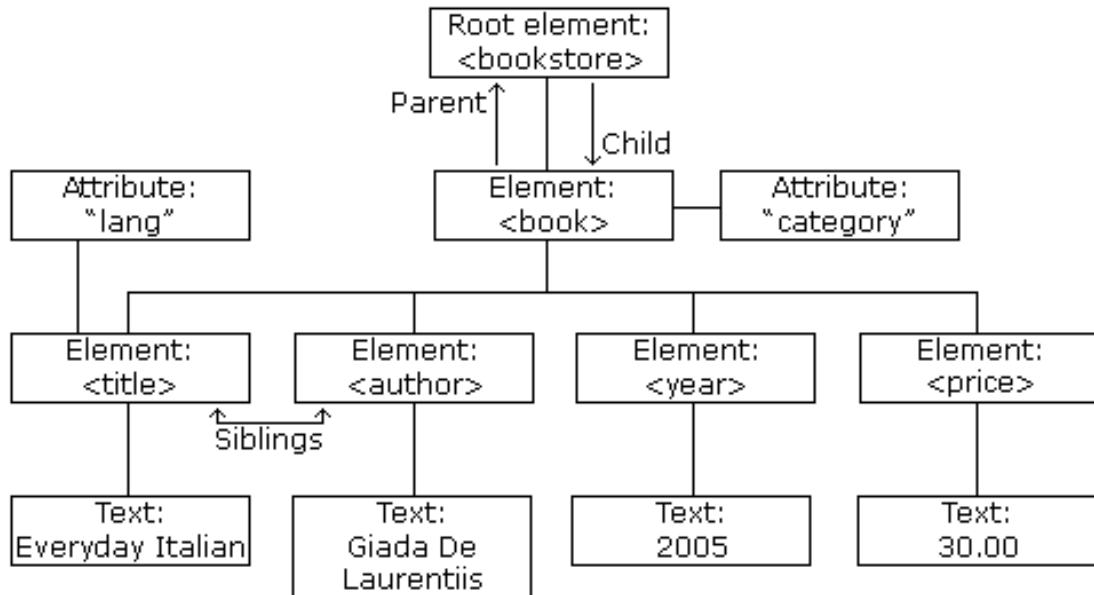
takse konteksti sobimatud analüüsivariandid. Analüüsi ülesandeks on leida lause struktuur, kas fraasistruktuur (millistest fraasidest lause koosneb) või sõltuvusstruktuur (kuidas sõnad lauses üksteisest sõltuvad). (Müürisep, 2000, 13, 17). Semantilise märgenduse käigus lisatakse igale sõnale teavet selle kohta missugusesse semantilisse klassi kuulutakse. Üldjuhul mõeldakse semantilise märgendamise all sõnatähenduste ühestamist. Oluline see, et korpus oleks korralikult ja standardiselt märgendatud, sest nii saab seda korpust kasutada erinevate eesmärkide tarbeks. Morfoloogiline analüüs on kõikide teiste märgendustasemete (süntaktiline, semantiline jne) alus. (Muischnek et al., 2003, 13–14)

3.4. 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 juurelementist (*root*), millel on alamelementid 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, allaval on ülem ja kolleegid on samal tasemel paiknevad alluvad. Kõikidel elementidel võib olla sisu (*text content*) ja atribuut ehk tunnus. (*XML Tutorial*, 2016) Joonis 1 illustreerib raamatupoe elementide ehk raamatute hierarhilist struktuuri:

Joonise 1 kujutamine XML-kujul (*XML Tutorial*, 2016):

```
<?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>
```



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

```

<book category="web">
    <title lang="en">Learning XML</title>
    <author>Erik T. Ray</author>
    <year>2003</year>
    <price>39.95</price>
</book>
</bookstore>
    
```

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 kooderingu. 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 väärthus oleneb raamatu valdkonnast. Elemandi *<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 väärtsuseks 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 eeldefinieeritud 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 reeglitelte:

1. eksisteerib juurelement;
2. elementidel peab olema lõpumärgend;
3. elementide pesitsemine ehk üksteise sees paiknemine (*nesting*) on rangelt määratletud;
4. atribuutide väärised 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: dokumendiübi 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)

3.5. Sõnaliikide kitsaskohad

“Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele” (Heiki-Jaan Kaa-lep, Kadri Muischnek, Kaili Müürisep, Andriela Rääbis, Külli Habicht).

Kadri Muischnek ja Kadri Vider “Sõnaliigituse kitsaskohad eesti keele arvutianalüüs”

Miskit saaks ka siit:

Rudolf Karelson “Taas probleemidest sõnaliigi määramisel”

Tiit Hennoste artikkel “Suulise kõne uurimine ja sõnaliigi probleemid” või “Suuline kõne ja morfoloogiaanalüsaator”

Üks huvitav artikkel Reili Arguselt on “Imitatiiivide kohast lastekeeles: reduplikatsioonist, morfoloogiast ja sõnaliigilisest ambivalentsusest”, kus räägitakse imitatiiividest ja onomatopoeetilistest sõnadest. Ja see teema haakub väga palju praktilise osaga: morfanalüüsatori kategoriad ei sobi hästi selliste sõnade jaoks.

AGA ma kindlasti ootan ettepanekuid!

4. CHILDES ja eesti keele alamkorpused

Selles peatükis tutvustatakse CHILDES-i korpust ja eesti keele alamkorpuseid, mida siinkirjutaja kasutab tulevase magistritöö materjalina. Lisaks antakse lühilevaade alamkorpuste standardiseerimise probleemidest.

4.1. CHILDES

CHILDES (*Child Language Data Exchange System*) on *Talkbanki* alamkorpus, mis loodi 1984. aastal Brian MacWhinney (Carnagie Mellon ülikool) ja Catherine Snow (Harvardi ülikool) poolt selleks, et koondada kokku erinevate kleeleuurijate kogutud keelematerjali eesmärgiga, et need oleksid kõigile vabalt kätesaadavad 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 funksioneerib *repositoriumina*. Repozitoriumist võib mõelda kui laost või arhiivist, kuhu üles laetud materjali talletatakse digitaalselt. Repozitoriumi jaoks on oluline, et korpused oleksid avalikult kätesaadavad 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 kategoorialasse:

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

Andmebaasi korpuste tekstide/lindistuste transkribeerimiseks/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 baasta-

semel. Lisaks CHAT-käsiraamatule on keeleuurijatel võimalus kasutada ka analüüsistarkvara ja redaktorit CLAN (*Computerized Language Analysis*), mis abistab keeleuurijat korpuse transkribeerimisel, kodeerimisel ja analüüsimisel. 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: germani keeled (afrikaani, taani, hollandi, inglise, saksa, norra ja rootsi keel), romani keeled (katalaani, prantsuse, itaalia, portugali, rumeenia ja hispaania keel), slaavi keeled (horvaadi, vene, serbia ja sloveenia keel), keldi keeled (iiri ja kõmri keel), afroasia keeled (araabia, heebrea ja berberi keeled), hiina-tiibeti keeled (mandariini hiina, kantoniini, taiwani ja tai keel), draviidi keeled (tamili keel), uurali keeled (eesti ja ungari keel), indoiraani keeled (farsi keel), austroneesia keeled (indoneesia keel), kreeka, cree ja baski keeled. 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 enam on esindatud inglise, saksa ja prantsuse keel. (Gillis, 2014, 2–5)

Transkriptsioonid algavad päisega (ingl k. *header*), kus antakse informatsiooni linnistuse 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 kommentaaririda. (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)

4.2. Mida teistes keeltes tehtud vms...

Siin rääkida ehk sellest, et kas ja millised teised keeled on CHILDES-is morfanalüüsiga esindatud, nt heebrea keeles on sellega usinasti tegeldud. Aga ma ei ole kindel, et see alapeatükk siia vahele sobib.

4.3. Alamkorpuste standardiseerimise probleemid

TO-DO:

SIIN TAHAKS KIRJUTADA KA “From CHILDES to TalkBank” (Brian MacWhinney, CHILDES-i asutaja), sest siin tuuakse välja üldised transkriptsiooni probleemid.

SIIN tahaks puudutada ka morfoloogiliste vigade teemat (et mida veaks pidada, vea liigitusprobleemist, vigade transkribeerimis- ja kodeerimisprobleemidest, vea-kodeerimisvõimalustest ja see kõik puudutab CHILDES-it). Aluseks oleks Arguse artikkel “Eesti lastekeelekorpuse morfoloogiliste vigade märgendamisest ja liigitamisest”.

END-TO-DO

Reili Argus kirjeldab oma artiklis (Argus, 2007) mõningaid transkribeerimise ja CHILDES-i tarkvara kasutamisega seonduvaid probleeme.

Esiteks, CLAN-i analüüsitarvvara on mõeldud inglise keelele, seega tuleb eesti keele analüüsimal arvestada sellega, et eesti keel on võrreldes inglise keelega süntetilisema süsteemiga keel. Seega, kui keeleuurija tahab CLAN-i tarkvara 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 väärata arvude tekkimist. Homonüümide eristamist tuleb teha näiteks käsitsi.(Argus, 2007, 70)

Argus väidab, et kuna lindistuste transkribeerimisel kasutatakse kuuldeortograafiat, siis need transkriptsioonid ei anna töetruud pilti sellest, milline on lapse tegelik keelekasutus. Kui juba suulise kõne automaatne analüüsime on keeruline, siis on lapse suulise keele analüüsime veel keerukam. Lindistuste puhul on tegemist spontaanse suulise kõnega, mis sisaldab elemente, mida pole tarvis analüüsida, nt häältsused. 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 tekib 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-süsteem kohtleb seda kui korduvat üksust. Lisaks sõnakordusele on probleemiks ka onomatopoeetilised sõnad, mida esineb lapsekeeltes 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 lapsekeeltes 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 järel 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 = järelle

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 analüüsitarvara ei võimalda eesti keele morfoloogilist analüüsimist ja selle käsitse tegemine oleks väga ajamahukas töö. Seega, praegusel hetkel on lapsekeele uurijatel automaatse statistika tegemine raskendatud ja paraku tehakse distribuutsionianalüüse käsitsi (Argus, 2007, 78).

4.4. Eesti keele alamkorpused

CHILDES-i andmebaasis on eesti laste suulise kõne lindistused olnud alates 1998. aastast. 2016. aasta märtsikuu seisuga koosneb eesti lastekelle korpus seitsmest alamkorpusest, mis on oma nimed saanud korpuse koostajate järgi: Argus, Beek, Kapanen, Kohler, Kõrgesaar, Vija ja Zupping (CHILDES, 2016). Tabelid 1–7 kajastavad laste vanuselist jaotumist alamkorpuste lõikes. Lisaks on välja toodud ka lapse nimi ja sugu, lapsega tehtud sessioonide arv, lapse ja lapsele suunatud kõne ehk hoidjakeele sõnade arv igas sessioonis.

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

Tabel 1: Vija korpus

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

Tabel 2: Arguse korpus

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	1841	3484
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	90165	155984

Tabel 3: Kõrgesaare korpus

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

Tabel 4: Beeki korpus

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

Tabel 5: Kapaneni korpus

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

Tabel 6: Zuppingi korpus

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

Tabel 7: Kohleri korpus

Korpus	Sõnade arv	% kogu korpusest
Vija	226311	37%
Kõrgesaar	155984	25%
Argus	13373	2%
Beek	59136	10%
Kapanen	45675	7%
Zupping	42461	7%
Kohler	70712	12%
KOKKU	613652	100%

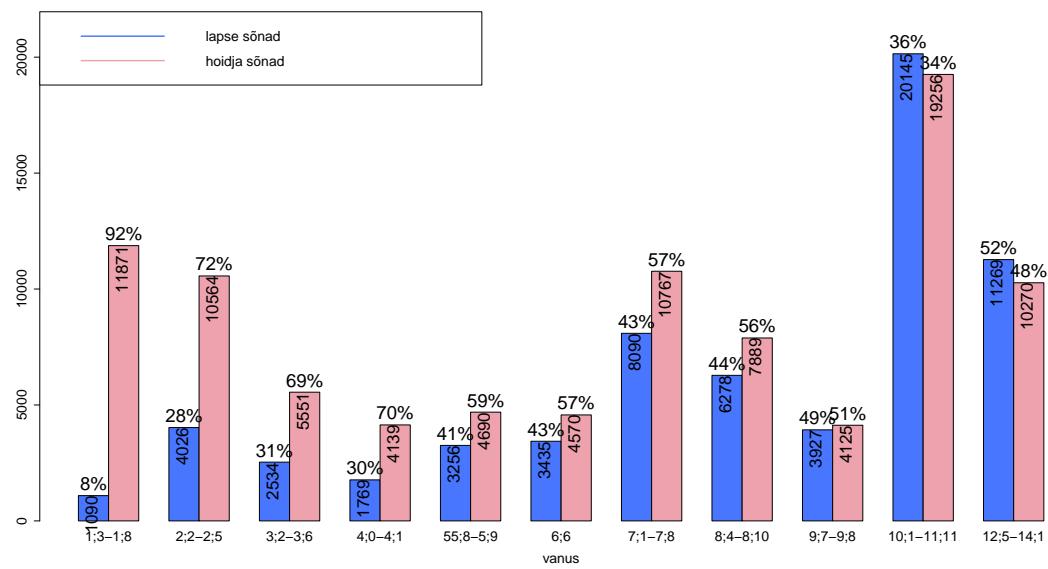
Tabel 8: Alamkorpuste % kogu korpuses

Kõige mahukamat on Vija, Kõrgesaare ja Kohleri korpused. Neist mahukaim on Vija korpus, moodustades kogu korpusest 37%. Korpus sisaldab 226311 sõna, milles 110381 olid lapse sõnad ja 115930 hoidjasõnad. Lindistusi tehti Andreadsega vahemikus 1;7–3;1 eluaastat. Kõrgesaare korpus moodustab kogu korpusest 25% ja koosneb 155984 sõnast, neist 65819 on lapse sõnad ja 90165 hoidjasõnad. Materjal pärineb lindistustest 12 erineva lapsega vahemikus 1;3–14;1 eluaastat. Siia pole sisestatud transkriptsioone vestlustest, mille osalejateks olid vaid täiskasvanud. Kohleri korpus moodustab kogu korpusest 12% ja sisaldab 70712 sõna, lapse sõnade hulk 10053 ja hoidjakeele sõnade hulk 60659. Lindistusi tehti 8 erineva lapsega vahemikus 0;11–2;3 eluaastat.

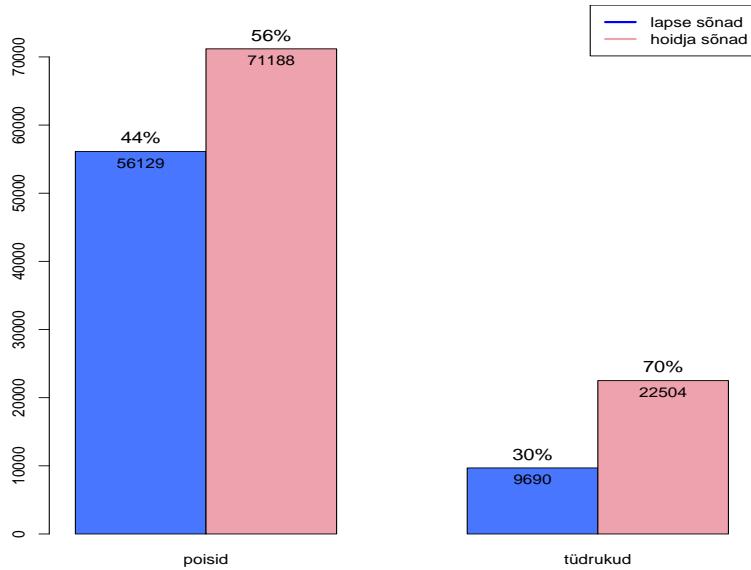
Beeki korpus moodustab kogu korpusest 10% ja sisaldab 59136 sõna, milles lapse sõnad on 10164 ja hoidjasõnad 48972. Lindistusi tehti Liisbetiga vahemikus 0;9–2;5 eluaastat. Kapaneni korpus moodustab kogu korpusest 7%, sisaldades 45675 sõna, neist 15938 lapse sõnad ja 29737 hoidjakeele sõnad. Kapaneni materjal pärineb lindistustest Martinaga vahemikus 1;3–2;7 eluaastat. Zuppingi korpus moodustab samuti kogu korpusest 7%, sisaldades 42461 sõna, neist 11632 on lapse sõnad ja 30829 hoidjakeele sõnad. Kõik lindistused on tehtud Lindaga vahemikus 1;3–4;2 eluaastat. Mahult kõige väiksem on Arguse korpus (2%). Korpus sisaldab 13373 sõna, milles 4220 on lapse ja 9153 hoidjasõnad. Lindistusi tehti Hendrikuga vahemikus 1;8–2;5 eluaastat. CHILDES-i eesti keele alamkorpuuse kogu suuruseks on 613652 sõna.

4.5. Tabelid

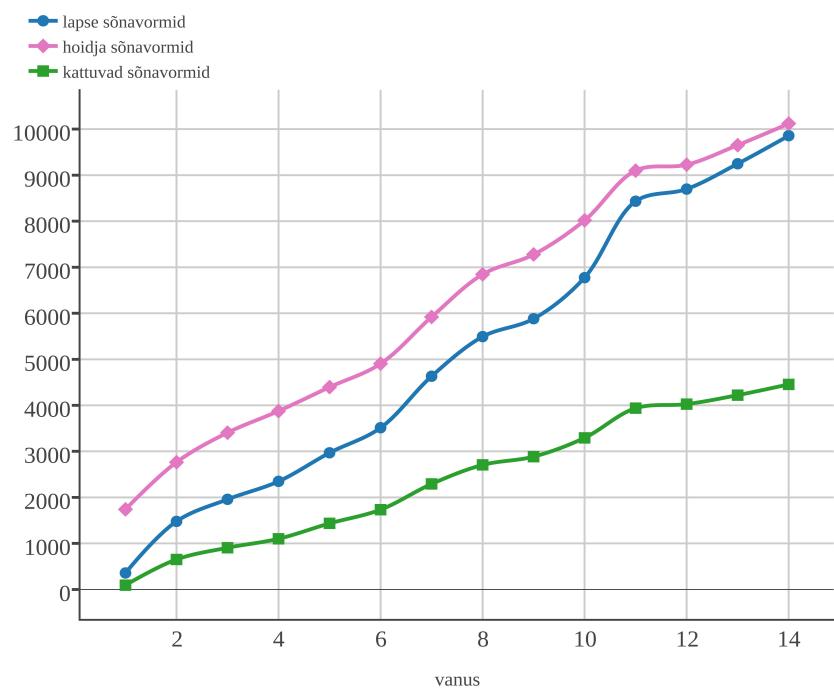
4.5.1. Kõrgesaar



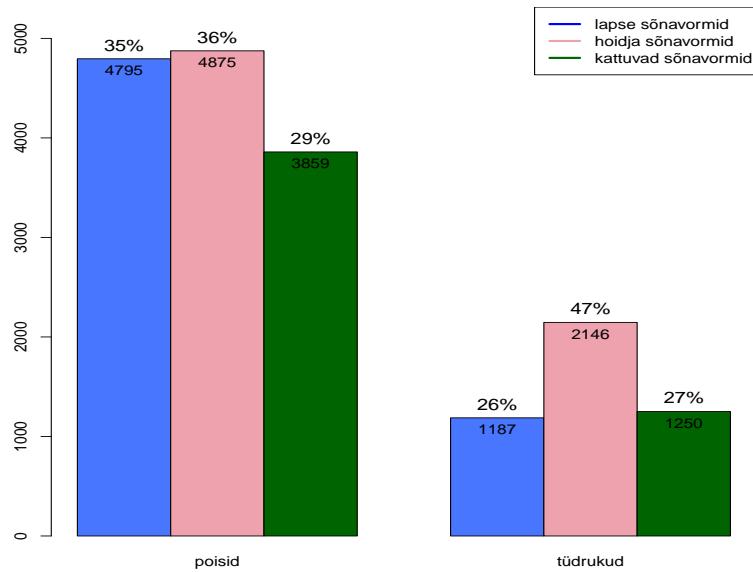
Joonis 2: Kõrgesaar: lapse ja hoidja sõnade vanuseline jaotumine



Joonis 3: Kõrgesaar: lapse ja hoidja sõnade sooline jaotumine

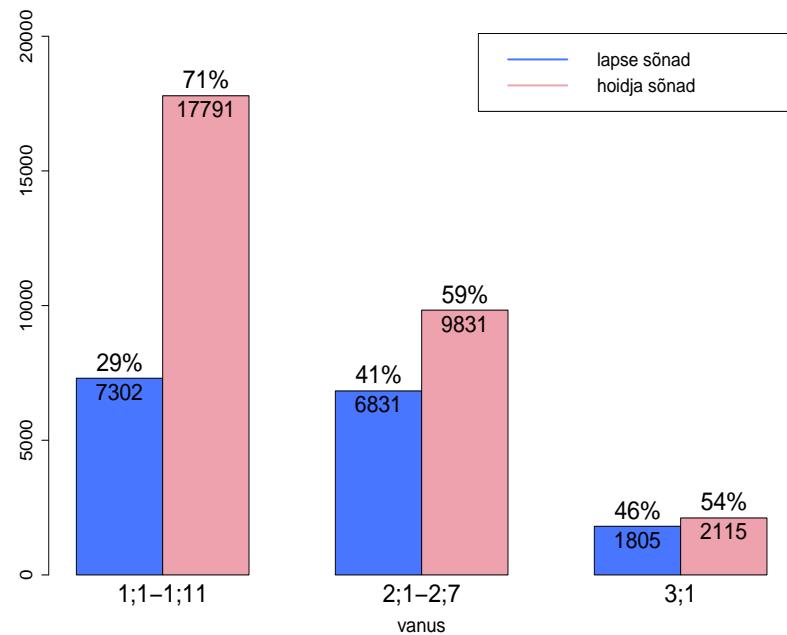


Joonis 4: Kõrgesaar: sõnavormide kasv vanuseliselt

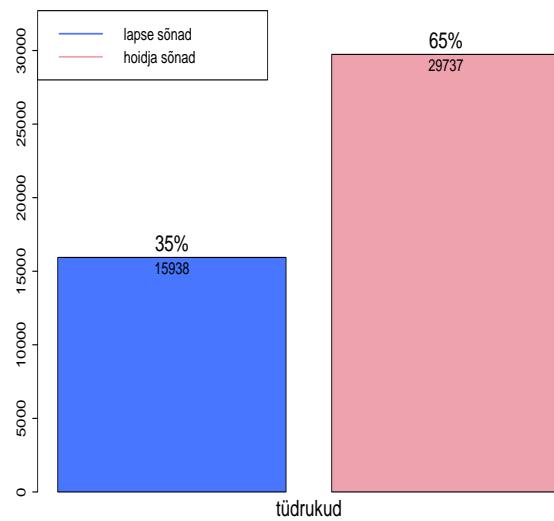


Joonis 5: Kõrgesaar: sõnavormide sooline jaotumine

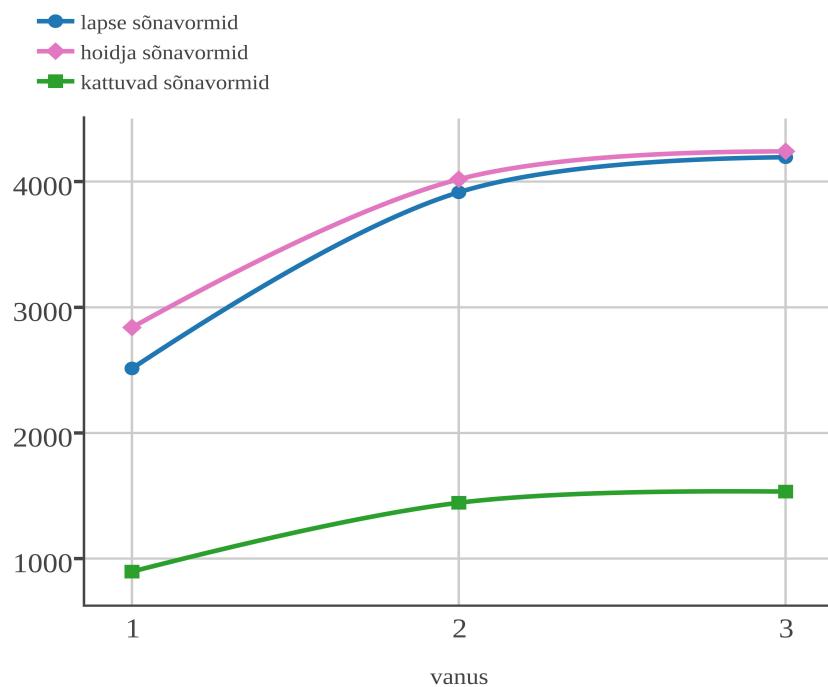
4.5.2. Kapanen



Joonis 6: Kapanen: lapse ja hoidja sõnade vanuseline jaotumine

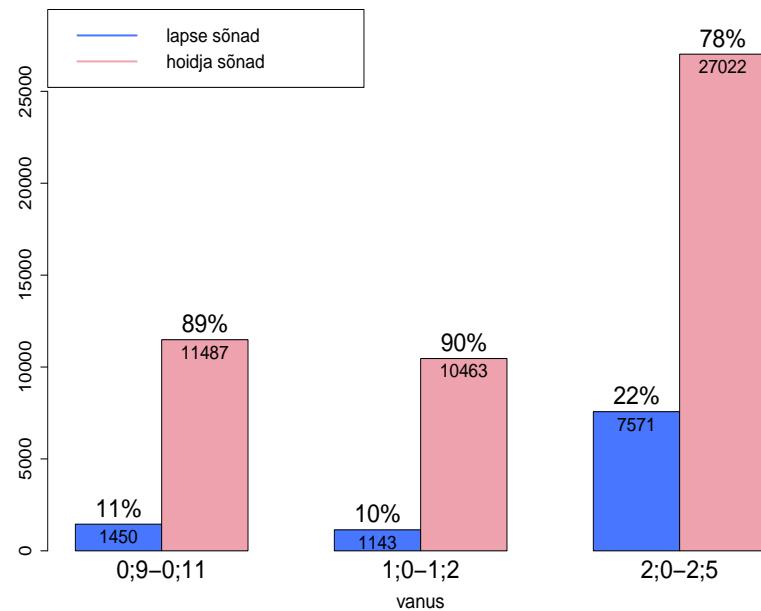


Joonis 7: Kapanen: lapse ja hoidja sõnade sooline jaotumine

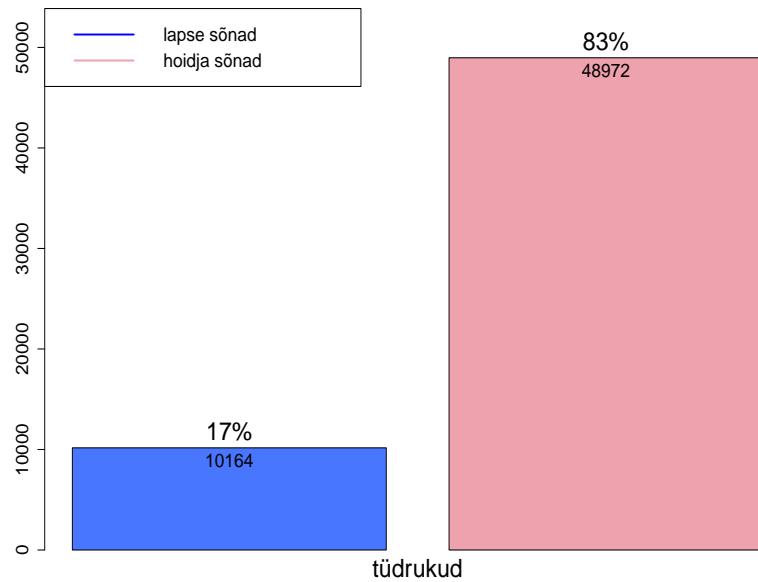


Joonis 8: Kapanen: sõnavormide kasv vanuseliselt

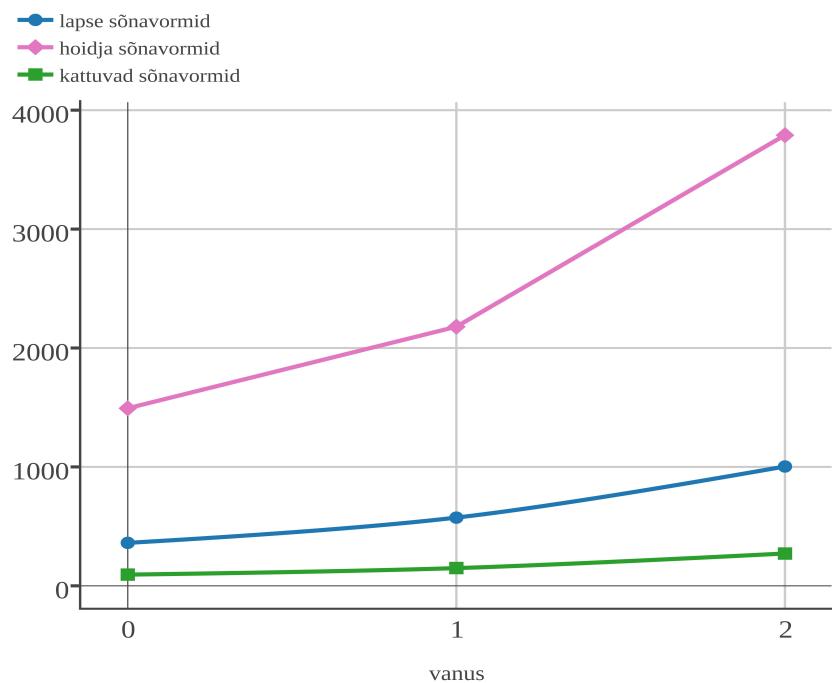
4.5.3. Beek



Joonis 9: Beek: lapse ja hoidja sõnade vanuseline jaotumine

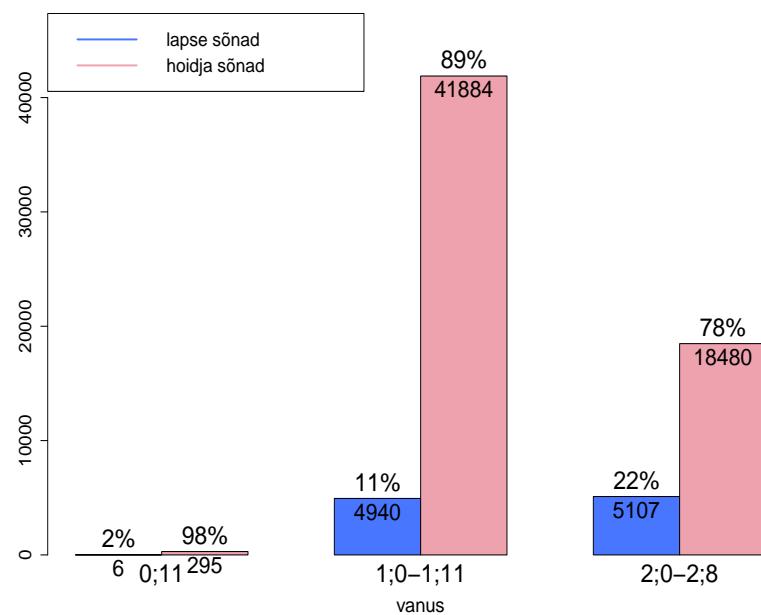


Joonis 10: Beek: lapse ja hoidja sõnade sooline jaotumine

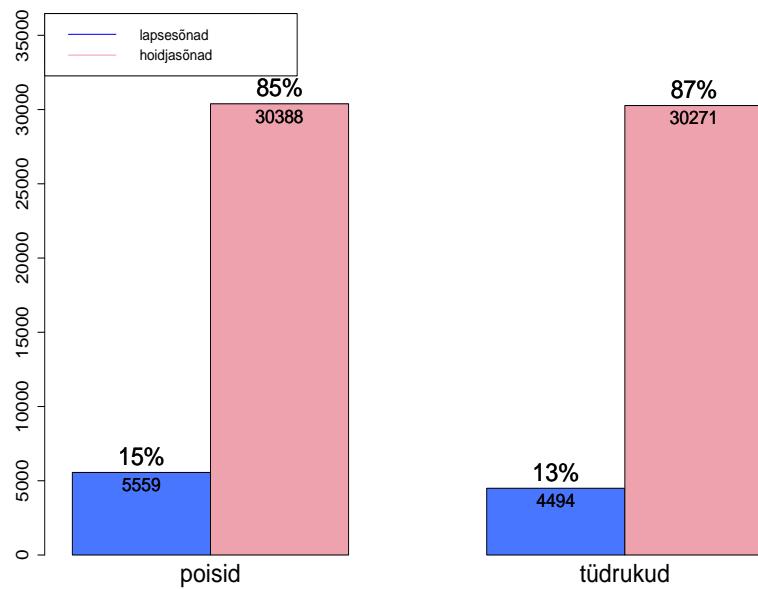


Joonis 11: Beek: sõnavormide kasv vanuseliselt

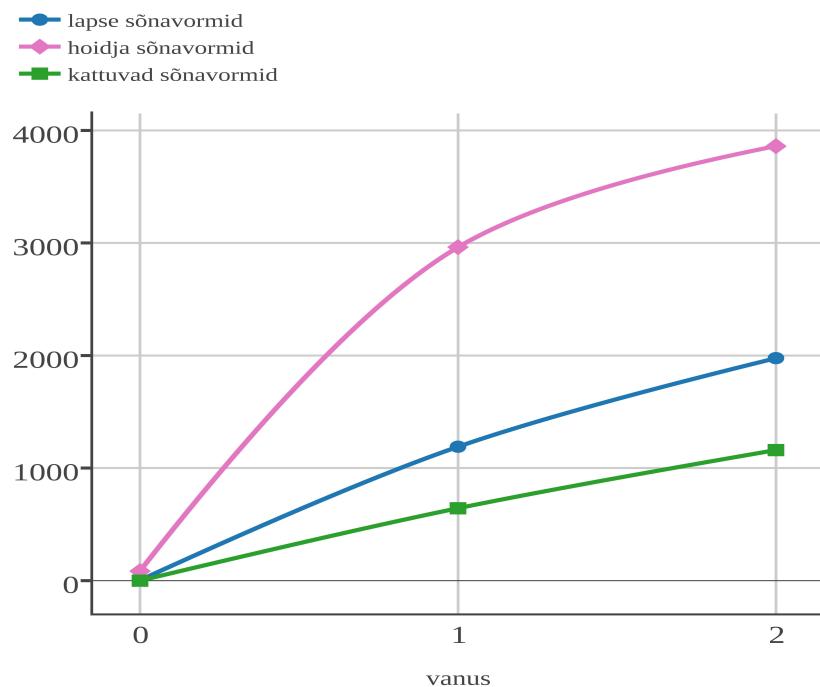
4.5.4. Kohler



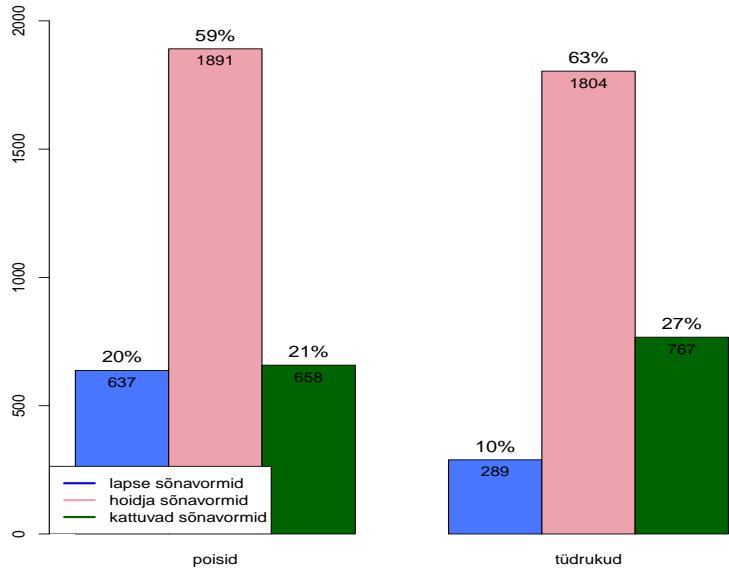
Joonis 12: Kohler: lapse ja hoidja sõnade vanuseline jaotumine



Joonis 13: Kohler: lapse ja hoidja sõnade sooline jaotumine

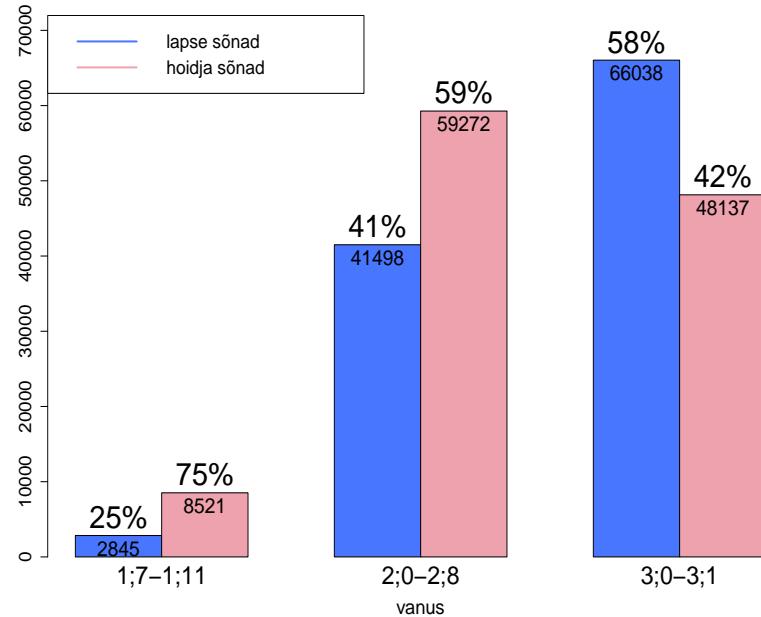


Joonis 14: Kohler: sõnavormide kasv vanuseliselt

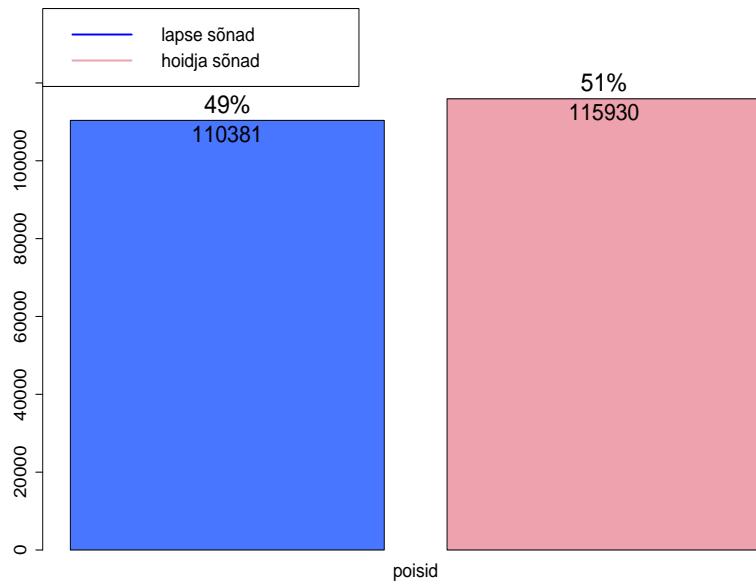


Joonis 15: Kohler: sõnavormide sooline jaotumine

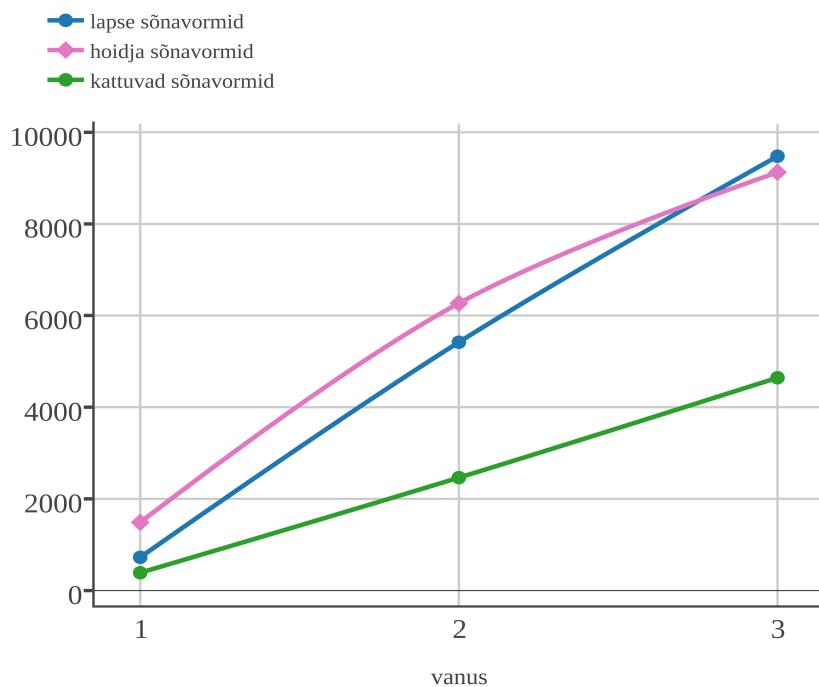
4.5.5. Vija



Joonis 16: Vija: lapse ja hoidja sõnade vanuseline jaotumine

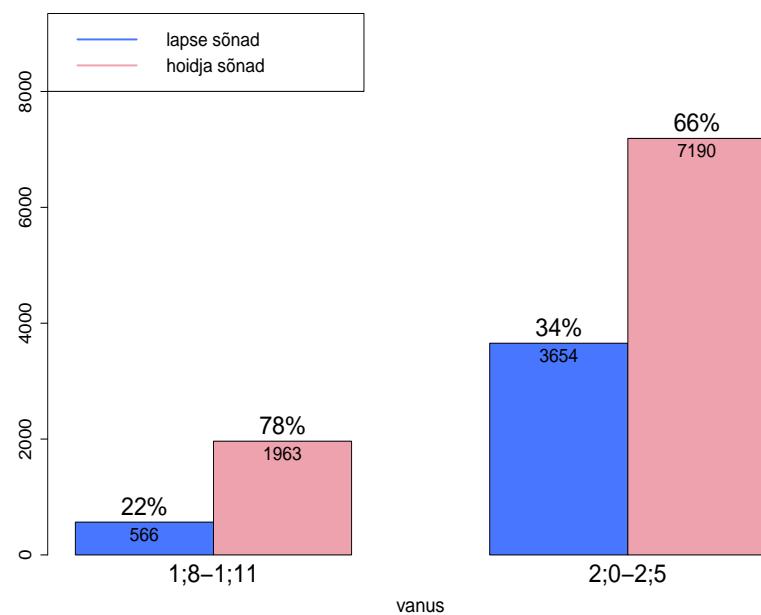


Joonis 17: Vija: lapse ja hoidja sõnade sooline jaotumine

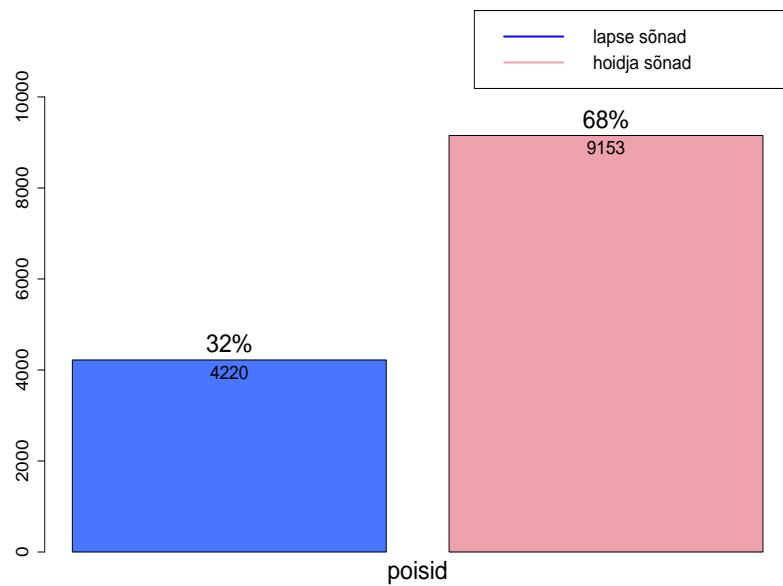


Joonis 18: Vija: sõnavormide kasv vanuseliselt

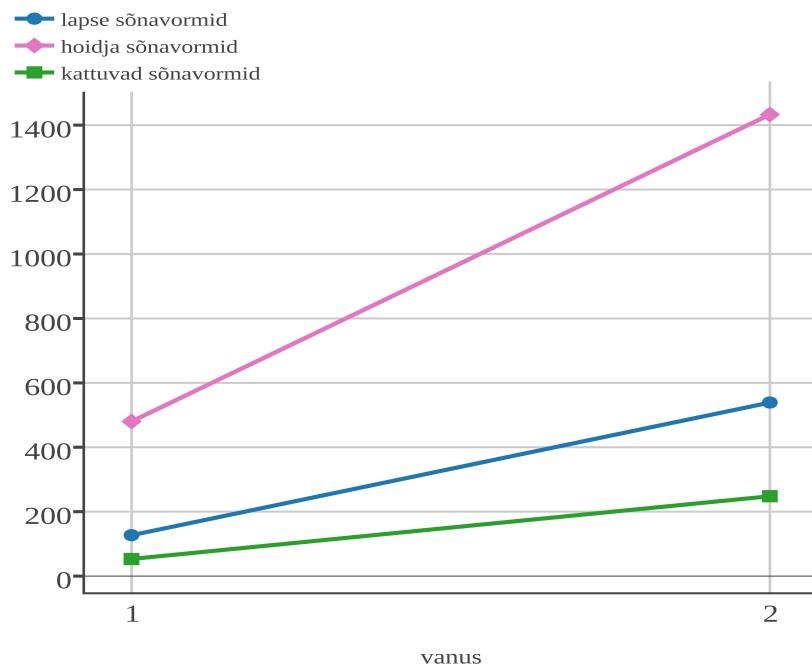
4.5.6. Argus



Joonis 19: Argus: lapse ja hoidja sõnade vanuseline jaotumine

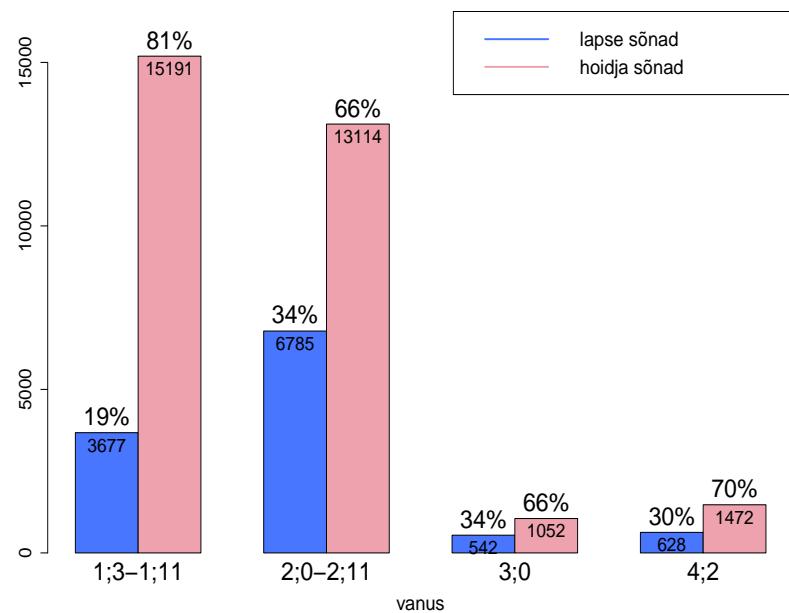


Joonis 20: Argus: lapse ja hoidja sõnade sooline jaotumine

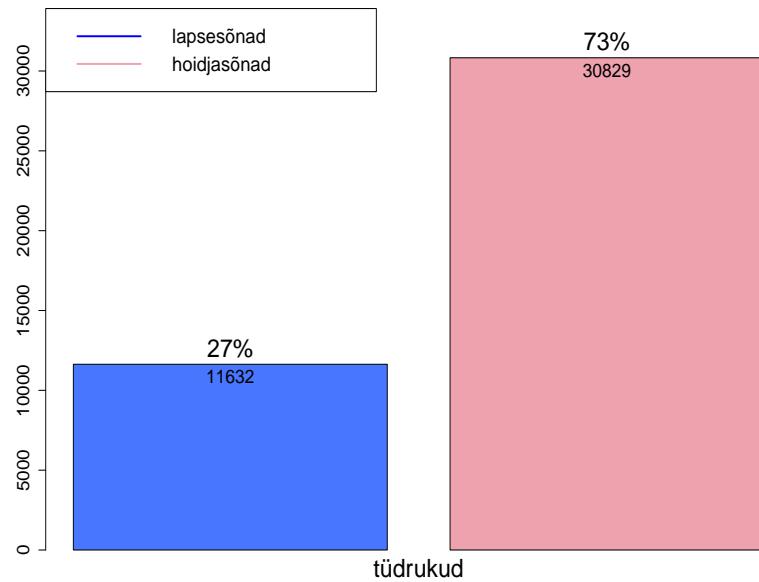


Joonis 21: Argus: sõnavormide kasv vanuseliselt

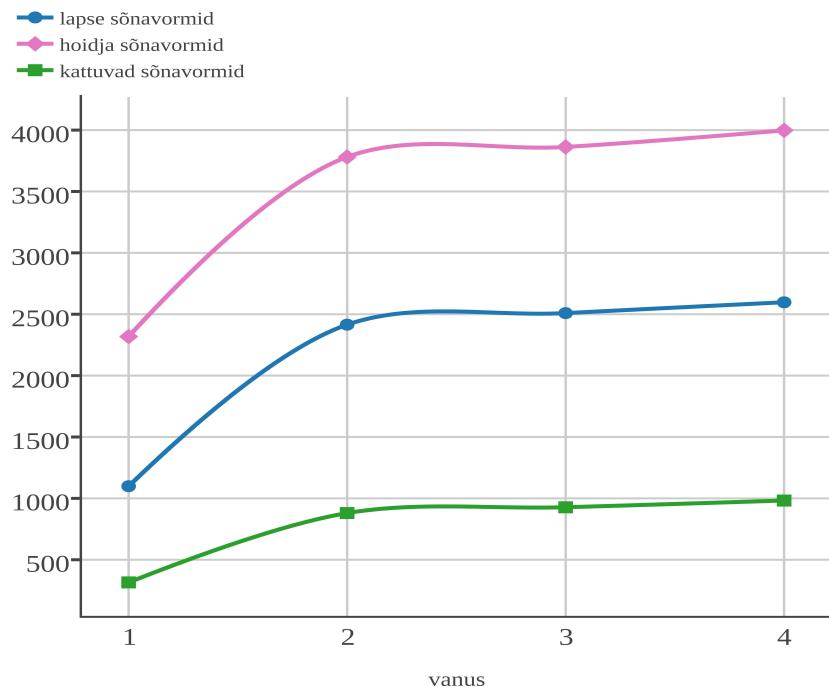
4.5.7. Zupping



Joonis 22: Zupping: lapse ja hoidja sõnade vanuseline jaotumine

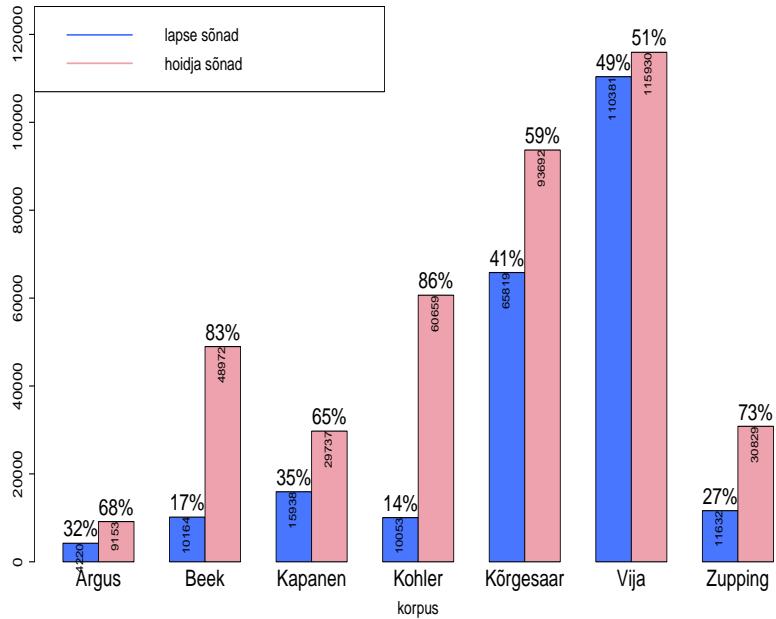


Joonis 23: Zupping: lapse ja hoidja sõnade sooline jaotumine



Joonis 24: Zupping: sõnavormide kasv vanuseliselt

4.5.8. Kõik korpused



Joonis 25: hoidja ja lapse sõnade jaotumine korpustes

Korpus	Lapse sõnavormid	Hoidja sõnavormid	Kattuvad sõnavormid	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 9: Sõnavormide jaotumine alamkorpustes

5. Morfoloogiliselt märgendatud lapsekeelete korpus

5.1. Tööprotsess

Magistritöö eesmärk on luua eesti morfoloogiliselt märgendatud lapsekeelete 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 CHAT-faile XML-kujule konverteerida, oleks tarvis, et kõik alamkorpused oleksid ühtsel kujul transkribeeritud ja kodeeritud. Peatükis 5.4 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. Põhjas seisneb selles, et CHAT-käsiraamatus on väga suur ja lai valik kodeeringuid, mida on paraku ühel inimesel raske hallata. Näide (5) illustreerib seda, kuidas juba üsna lühikeses transkriptsiooni lõigus võib kodeeringute kasutus väga mitmekesine olla (kodeeringu seletus paikneb lausungi järel).

(5)

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

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

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

*MOT: see ei ole kefir ju .

*CHI: kefir . [+ sr] | *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)

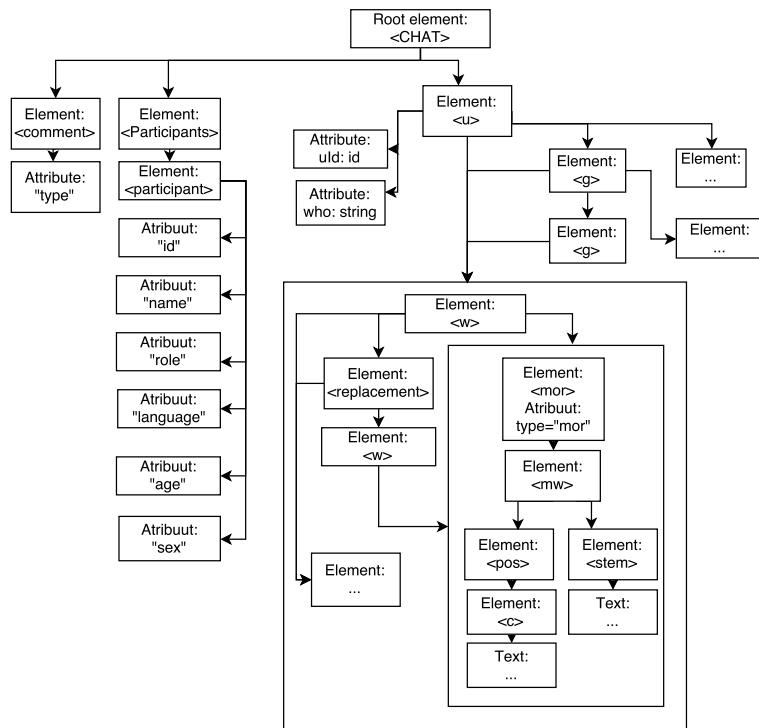
On arusaadav, et iga uurija transkribeerib ja kodeerib lindistusi lähtuvalt enda eesmärkidest. Ühelt poolt on hea, et CHAT-käsiraamatus on niivõrd detailne kodeering, kuid teisalt võib selles orienteerumine vägagi raskeks osutuda. Kuna sellise konverteerimisskripti kirjutamise töömaht oleks selle magistritöö 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 tuleb nendele XML-failidele ka lisada morfoloogiline tasand, mida neis failides ei ole. Selleks oli tarvis lähemalt tutvuda Talkbanki XML-skeema süntaksiga.

On arusaadav, et iga uurija transkribeerib ja kodeerib lindistusi lähtuvalt enda eesmärkidest. Ühelt poolt on hea, et CHAT-käsiraamatus on niivõrd detailne kodeering, 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 vaja-lik keelematerjal pärineb samuti CHILDES-i andmebaasist, kuid need on juba eel-nevalt CHAT-kujult XML-kujule konverteeritud failid. Aga kuna selle töö eesmärk on luua morfoloogiliselt märgendatud korpus, siis tuleb nendele XML-failidele ka lisada morfoloogiline tasand, mida neis failides ei ole. Selleks oli tarvis lähemalt tutvuda *Talkbanki* XML-skeema süntaksiga.

5.1.1. *Talkbanki* skeema

Joonisel 2 on kujutatud minu töö seisukohalt olulisimad skeema elemendid.



Joonis 26: Talkbanki skeema elemendid

Talkbanki skeema juurelement on *<CHAT>*, mille alluvad on *<comment>*, *<Participants>* ja *<u>*. Eleindi *<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 CHAT-käsiraamatu järgi kodeeritakse [: text] abil (vt näide (5) kifir [: kefir]), vt näide (7).

Näide (7) (Kohler; car030900.xml)

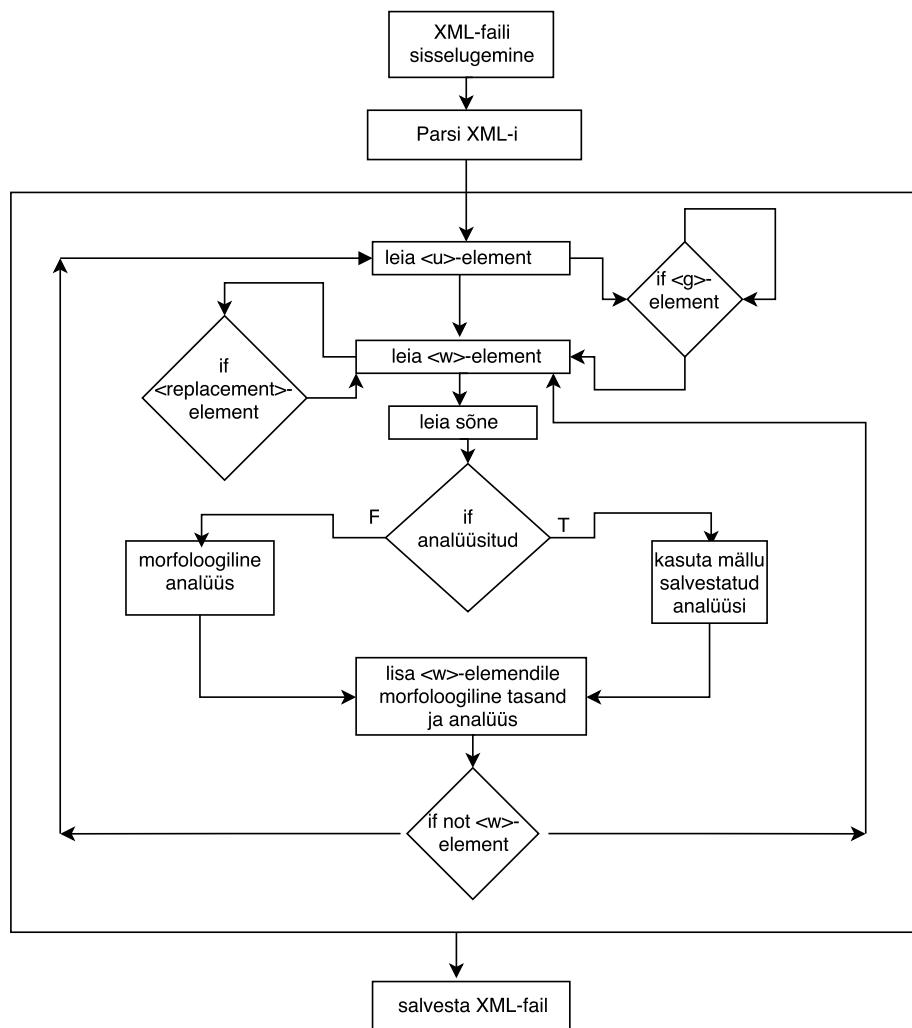
```

<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 *<mw>*, mille alluv on *<mw>* ehk *morphemic Word Type*. See jaguneb elemendiks *<pos>* ja *<stem>*. *<pos>* tähistab sõnaliiki (ingl k. *part of speech*). Selle alluvaks on *<c>* ehk sõna morfoloogiline kategoria, mille sisuks on mittetühi string. Element *<stem>* tähistab sõnatüve, mille sisuks on samuti mittetühi string.

5.2. Morfoloogilise info lisamise protsess



Joonis 27: Töövoog

Programmi kirjutamiseks on kasutatud Python 3.4 versiooni. Töövoog (vt joonis 3) 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, aga 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 kulkas protsess. Seejärel lisatakse igale sõnele morfoloogiline tasand. Programm genereerib kirjutatud koodis jätk-järgult puu elemendid ning morfoloogilisele tasandile jõudes hakkab sõne analüüse neile vastavatesse elementidesse lisama (vt joonis 2 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>
  ...
  ...
```

6. Kokkuvõte

Seminaritöös kirjeldati korpuse olemust ja selle tähtsust keeleuurijale. Korpuse mõte on seisneb selles, et sealt võimalikult lihtsalt olulist infot kätte saada, aga kahjuks korpuste standardiseerimine ja loomine pole nii lihtne töö. Kõik algab algmaterjalist.

Selle töö eesmärk oli lühidalt tutvustada ja näidata, et selleks, et lapsekeele korpust luua, tuleks transkriptsioonides esmalt selgeks teha, et mida ja kuidas märgendada. Kui see on selgeks tehtud, siis tuleks dialoogide transkribeerimisel sellest ka kinni pidada ja teha seda järjepidevalt. Eelmises peatükis kirjeldasin vaid mõningaid transkriptsioonidega seotud probleeme. Paraku on nii, et see esimene tase (transkriptsioon) mõjutab oluliselt vahepealseid tasandeid (standardiseerimine ja morfoloogilise info lisamine), mis omakorda mõjutavad lõpliku morfoloogiliselt märgendatud korpuse kvaliteeti.

Kasutatud kirjandus

- Argus, R. (2007). Eesti lastekeelekorpuse morfoloogilisest märgendamisest, *Tal-linna ülikooli keelekorpuste optimaalsus, töötlemine ja kasutamine* pp. 65–86.
- Chatter tarkvara (2016). <http://talkbank.org/software/chatter.html>. 18.04.2016.
- CHILDES (2016). Childe-i andmebaas, <http://childepsy.cmu.edu/data/>. 06.05.2015.
- eTenTen (2015). etenten, <http://www2.keelevaab.ee/dict/corpus/ettenten/about.html>. 05.03.2015.
- Gillis, S. (2014). *Child Language Data Exchange System*, pp. 74–78.
- 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>. 05.03.2015.
- MacWhinney, B. (2016). Part 1: The chat transcription format. the childe project: Tools for analyzing talk – electronic edition, <http://childepsy.cmu.edu/manuals/CHAT.pdf>. 18.04.2016.
- Muischnek, K. (2015a). Keelekorpused – sama mitmekesised kui keel ise, *Oma Keel 1*(30): 37–44.
- Muischnek, K. (2015b). Keeleressursid. "keeletehnoloogiaaine. 05.05.2015.
- Muischnek, K., Orav, H., Kaalep, H. & Ōim, H. (2003). Eesti keele tehnoloogilised ressursid ja vahendid. arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara, pp. 1–86.
- Müüriseep, K. (2000). *Eesti keele arvutigrammatika: süntaks*, doktoritöö, Tartu Ülikool.
- XML Tutorial (2016). <http://www.w3schools.com/xml/default.asp>. 18.04.2016.