TARTU ÜLIKOOL
HUMANTITAARTEADUSTE 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

SEE KOKKUVÕTE OLI SEMINARITÖÖ JAOKS. SEEGA, SEE POLE SEE "PÄRIS" KOKKUVÕTE. JA PEALKIRJAD ON KA ALLES ÜSNA ALGELISED.

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

Töö koosneb kahest peatükist. Esimeses peatükis räägitakse lähemalt korpusest, selle liigitusvõimalustest 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 lapsekeele korpuse loomise probleemidest. Tutvustatakse CHILDES-i korpust 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 formatiive (tunnused ja lõpud) 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 reeglipärasem 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 dešifreerimisega ehk püütakse öeldule tähendust omistada. Morfeemide hulk keeles on väga suur, seega pole laps võimeline korraga kõiki morfeeme ära õppima ning selle tõttu 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 hääldada 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 lõpud) osa. Reduplikatiivsed sõnad hõlbustavad morfoloogia omandamist, 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 tekitatakse kahesilbiline 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 analüüsimist. Sel perioodil esineb laste kõnes kõige rohkem tunnusteta kahesilbilisi sõnu. Pikemates vormides, kas loobutakse rõhutust ehk viimasest silbist (nt *sitikas* > *siti*, *traktor* > *takku*/*takka*), *rebane* > *repa* või jäetakse käändelõpp ära ja kasutatakse selle asemel genitiivi tüve (nt *poti* > *potile*). Kahesilbiliste vormide kasutust võib vaadata nii eelistuse kui ka piiranguna. Piirang selles mõttes, et laps ei pruugi olla suuteline hääldama pikemates sõnades teatud rõhutut silpi. Eelistus selles mõttes, et laps eelistab kahesilbilisi sõnu, kuna need on sisendkeeles kõige sagedasemad. Nii eesti kui ka soome keeles on lapsekeele uurijad märganud, et on perioode, mil laps pikendab ühesilbilise sõna kahesilbiliseks, nt *lutt* > *luti*, *pai* > *paia*. Näiteid on ka sellest, et liitsõnade omandamisel püütakse eelistada kahesilbilisi sõnu, nt *jõuluvana* > *jõvvu*. Vahetult enne protomorfoloogilist etappi lõpeb kahesilbiliste 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, partitiivi lõpp -*t* jääb hääldamata (*küpsist* > *kispi*), mitmuse nominatiivi lõpp -*d* jääb hääldamata (*mammud* > *mammu*), inessiivi lõpp -*s* jääb hääldamata (*vannis* > *vanni*, *aias* > *aia*). Konsonandi hääldamine sõltub paljuski konsonandi positsioonist: sõnaalguline konsonant jääb 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 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. (Argus 2008a: 23) VIIDE

Lõpuvaheldusega seotud vigu esineb üleminekul premorfoloogiliselt perioodilt protomorfoloogilisele perioodile. Enim valmistab probleeme II-välteliste konsonantlõpuliste 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 järgmiste 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 lõpufoneemide 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 lõpuvaheldusmalle 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 astmevahelduseta muuttüüpi. Näiteks *kiisu* ja *kutsu* saavad selle nihke tõttu partitiivi läbipaistva lõpu -*t*: *kiisu-t*, *kutsu-t*. Siinkohal on hõlbustav tegur ka vormihomonüümia, sest tuletatud sõnad nihkuvad sellisesse muuttüüpi, 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 **muna-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 konstrueerivad 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 peidetud 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 peidetud 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 traditsioonilist 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 keelest 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 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: *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, 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. 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 *eTenTen*, mis koosneb on 270 miljonist sõnast 686000 veebilehelt. (eTenTen, 2015)

Paralleelkorpus on korpus, mis sisaldab sama teksti vähemalt kahes keeles, mille üksused on paralleelistatud. *Paralleelistamine* tähendab paralleelteksti üksteise 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 parlamendidebattidest. 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 tõlkeid teistesse Läänemere-äärsetesse keeltesse. Keeletehnoloogia seisukohalt on paralleelkorpus väga oluline keeleressurss, nt saab paralleelkorpusi 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 lihtsasti kättesaadav. Selleks, et korpus ei jääks lihtsalt elektrooniliste tekstide arhiiviks, oleks tarvis korpusesse esmalt interpretatiivset infot lisada. Seda protsessi nimetatakse *korpuse märgendamiseks*. Korpuse märgendamise etapis lisatakse tekstidele 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 kombineerides (poolautomaatselt). 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). 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 (käändsõnal arv ja kääne, 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 analüüsivariandid 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 standardselt 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)

4. Sõnaliikide kitsaskohad

"Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele" (Heiki-Jaan Kaalep, Kadri Muischnek, Kaili Müürisep, Andriela Rääbis, Külli Habicht).

Kadri Muischnek ja Kadri Vider "Sõnaliigituse kitsaskohad eesti keele arvutianalüüsis"
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 "Imitatiivide kohast lastekeeles: reduplikatsioonist, morfoloogiast ja sõnaliigilisest ambivalentsusest", kus räägitakse imitatiividest ja onomatopoeetilistest sõnadest. Ja see teema haakub väga palju praktilise osaga: morfanalüüsaatori kategooriad ei sobi hästi selliste sõnade jaoks.

AGA ma kindlasti ootan ettepanekuid!

5. Morfoloogiliselt märgendatud lapsekeele korpus

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

5.1. 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 pidada 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 omandamise korpused;
2. teise keele keele omandamise korpused;
3. kakskeelsuse korpused ja
4. kliiniliste probleemide uurimiseks mõeldud korpused. (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 standartsel 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 analüüsimistarkvara 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: germaani keeled (afrikaani, taani, hollandi, inglise, saksa, norra ja rootsi keel), romaani keeled (katalaani, prantsuse, itaalia, portugali, rumeenia ja hispaania keel), slaavi keeled (horvaadi, vene, serbia ja sloveenia keel), keldi keeled (iiri ja kõmri keel), afroaasia 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)

5.2. Eesti keele alamkorpused

CHILDES-i andmebaasis on eesti laste suulise kõne lindistused olnud alates 1998. aastast. 2016. aasta märtsikuu seisuga koosneb eesti lastekeele 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 infot iga alamkorpuse lindistuste arvu ja suuruse kohta. Lisaks on välja toodud ka laste nimed ning lapse vanuseline ja sooline jaotumine.

Lapse nimi	Vanus	Sugu	Sessioonid	Suurus
Andreas	1;7–1;11	m	7	126.9kB
	2;0–2;8		37	1100kB
	3;0–3;1		30	998.85kB
Kokku			74	2225.75kB

Tabel 1: Vija korpus

Lapse nimi	Vanus	Sugu	Sessioonid	Suurus
Hendrik	1;8–1;11	m	5	25.6kB
	2;0–2;5		12	103.5kB
Kokku			17	129.1kB

Tabel 2: Arguse korpus

Lapse nimi	Vanus	Sugu	Sessioonid	Suurus
Andri	11;7–11;9	m	2	63.6kB
Arabella	11;8	f	1	19kB
Artur	1;4	m	1	29.2kB
Gregory	6;6–10;5	m	10	344.7kB
Harley	4;0–14;1	m	19	387.8kB
	4;0	f	2	10.7kB
Hellyn	8;7	f	1	26.4kB
Jaana	2;5	f	1	25.1kB
Kaisa	5;8–5;9	f	2	59.9kB
Mia	2;3	f	1	49.2kB
Olivia	3;2	f	1	37.8kB
Ruuben	1;3–3;6	m	4	96.4kB
Sirlin	1;3	f	1	19.3kB
Kokku			46	1169.1kB

Tabel 3: Kõrgesaare korpus

Lapse nimi	Vanus	Sugu	Sessioonid	Suurus
Liisbet	0;9–0;11	f	6	103.7kB
	1;0–1;2		7	132.3kB
	2;0–2;5		7	212.2
Kokku			20	448.2kB

Tabel 4: Beeki korpus

Lapse nimi	Vanus	Sugu	Sessioonid	Suurus
Martina	1;3–2;1	f	2	78.4kB
n/a	2;3–2;7	n/a	2	62.5kB
	1;10–3;1	f	7	220.4kB
Kokku			11	361.3kB

Tabel 5: Kapaneni korpus

Lapse nimi	Vanus	Sugu	Sessioonid	Suurus
Linda	1;5–1;7	f	3	40.4kB
	2;0–2;9		3	64kB
n/a	1;3–1;11	f	6	80kB
	2;1–2;11		9	153.9kB
	3;0		1	20.3kB
	4;2		1	16.6kB
Kokku			23	375.2kB

Tabel 6: Zuppingi korpus

Lapse nimi	Vanus	Sugu	Sessioonid	Suurus
Anna	1;10–1;11	f	4	51.7kB
	2;0–2;1		3	42kB
Carlos	1;7–1;10	m	9	98.5kB
Helen	1;1–1;10	f	7	94.9kB
Henri	2;2–2;3	m	3	33.4kB
Mari	2;5–2;8	f	7	97.7kB
Sandor	1;2–1;10	m	7	120.3kB
	2;2		3	58.9kB
Stella	0;11	f	1	3.9kB
	1;0–1;6		8	82.2kB
Taimo	1;5–1;11	m	9	86.2kB
Kokku			61	769.7kB

Tabel 7: Kohleri korpus

Kõige mahukamad on Vija, Kõrgesaare ja Kohleri korpused. Neist mahukaim on

Vija korpus, koosnedes 74 transkriptsioonist. Lindistusi tehti Andreasega vahemikus 1;7–3;1 eluaastat. Kõrgesaare korpus koosneb 46 transkriptsioonist ja 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 koosneb 61 transkriptsioonist. Lindistusi tehti 8 erineva lapsega vahemikus 0;11–2;3 eluaastat.

Mahult kõige väiksemad on Arguse, Zuppingi, Kapaneni ja Beeki korpus. Arguse korpus koosneb 17 transkriptsioonist ja lindistusi tehti Hendrikuga vahemikus 1;8–2;5 eluaastat. Beeki korpus koosneb 20 trankriptsioonist ja lindistusi tehti Liisbetiga vahemikus 0;9–2;5 eluaastat. Kapaneni korpus sisaldab 11 trankstriptsiooni. Muist trankriptsioonides puudub info lapse nime ja soo kohta, kuid võib oletada, et tegu on Martinaga, ja lindistusi tehti vahemikus 1;3–2;7 eluaastat. Zuppingi korpus koosneb 23 transkiptsioonist. Ka siin mõnes failis on info puudulik, aga võib samuti oletada, et kõik lindistused on tehtud Lindaga vahemikus 1;3–4;2 eluaastat. CHILDES-i eesti keele alamkorpus koosneb 252 CHAT-failist ja ühtlasi on ka iga fail konverteeritud XML-kujule.

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 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)

5.3. 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.

5.4. 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, veakodeerimisvõimalustest 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 tranksribeerimise ja CHILDES-i tarkvara kasutamisega seonduvaid probleeme.

Esiteks, CLAN-i analüüsitarkvara 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 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 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-süsteem 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).

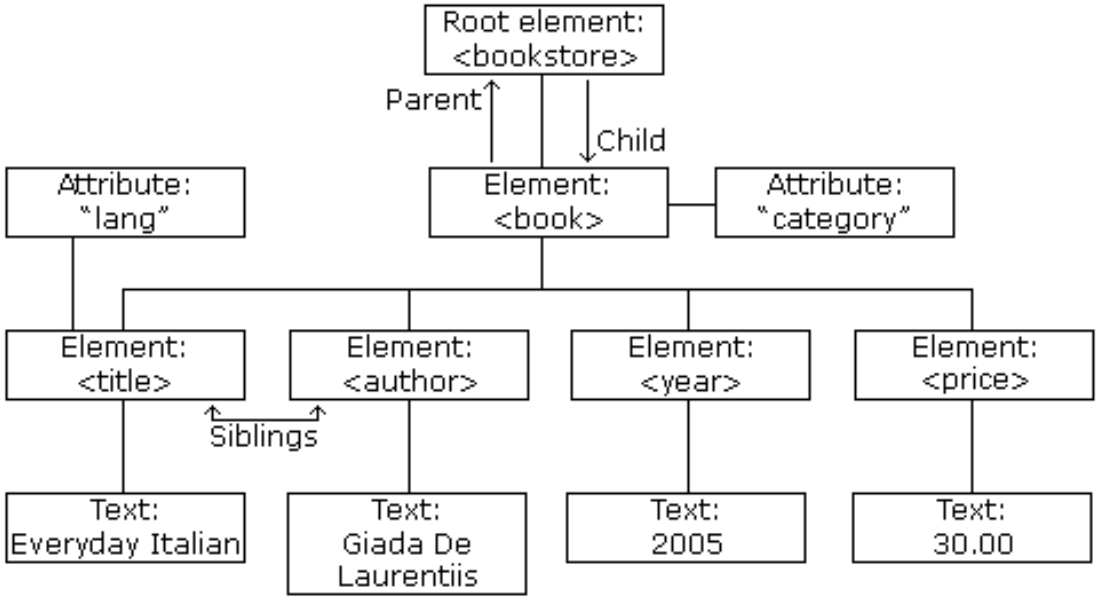
6. Praktiline osa

Siit edasi kirjutaks oma praktilisest osast.

6.1. Mis on XML?

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. (*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 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>
  <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 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 eeldefi-

neeritud 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 (*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)

6.2. Tööprotsess

Magistritöö eesmärk on luua eesti morfoloogiliselt märgendatud 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 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 standartsel kujul, siis oleks konverteerimisskripti tegemine muutunud väga keeruliseks ja ülejõukäivaks ülesandeks. Põhjus seisneb selles, et CHAT-käsiraamatus on väga suur ja lai valik kodeeringuid, mida on paraku ühel inimesel raske hallata, vt näide (5).

(5)

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

*MOT: kus sa **+/**.

*CHI: **+ <** (h)akkas põlema .

*MOT: see ei ole kefiir ju .

*CHI: kefiir . **[+** sr**]**

*MOT: see on piim .

*FAT: mis see kook teeb ?

*FAT: tuleb ära panna **[=** visata**]** või ?

*MOT: mina ei tea , vist jah .

*CHI: kuidas emme küpsetab saia , lihat**@n [*]** .

%err: lihat=liha $MOR

*FAT: saia ei ei küpseta .

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

*CHI: **+ <** saia .

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

%err: lihat=liha $MOR

*FAT: liha ka jah .

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

%err: lihat=liha $MOR

%act: MOT koorib sibulat

...

*CHI: Atu **[:** Andreas**]** sõi **+...**

*CHI: **+"** a .

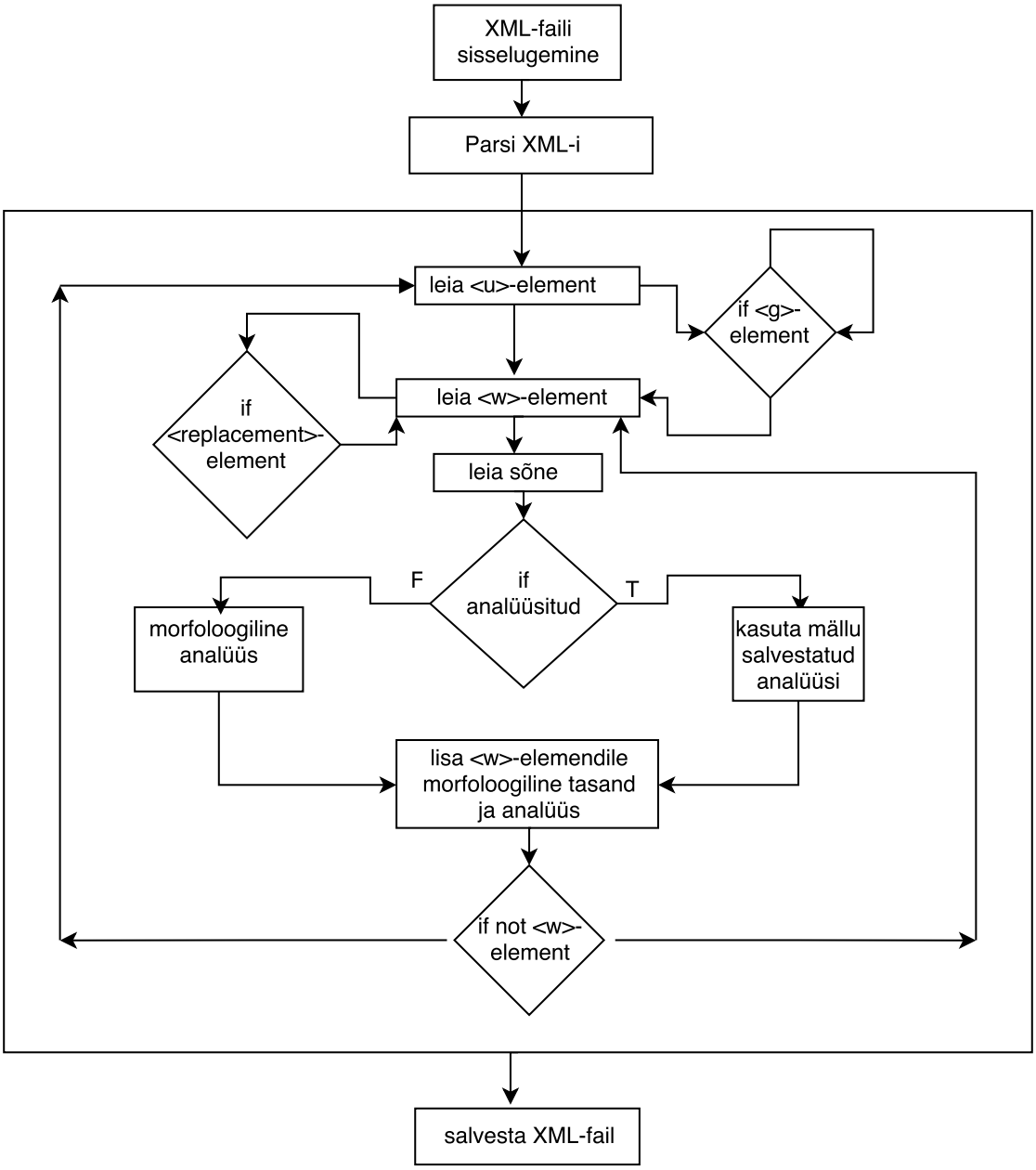
...

*CHI: käpad (h)aige **[*]** **[/]** käpad (h)aige **[*]** .

(Vija; 20018.cha)

Näites (5) on näha, kui mitmekesine võib ühes transkriptsioonis kodeeringute kasutus olla. On arusaadav, et iga uurija transkribeerib ja kodeerib lindistusi lähtuvalt enda eesmärkidest. Ühest käest 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 nendes failides ei ole. Selleks oli tarvis lähemalt tutvuda Talkbanki XML-skeema süntaksiga.

SIIA SKEEMA ELEMNTIDE KOHTA!!



Joonis 2: Töövoog

7. 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.

Viited

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

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

CHILDES (2016). Childes-i andmebaas, `http://childes.psy.cmu.edu/data/`. 06.05.2015.

eTenTen (2015). etenten, `http://www2.keeleveeb.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 childes project: Tools for analyzing talk – electronic edition, `http://childes.psy.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. "keeletehnoloogiaäine. 05.05.2015.

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

Müürisep, K. (2000). *Eesti keele arvutigrammatika: süntaks*, doktoritöö, Tartu Ülikool.

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