EESTI LASTEKEELEKORPUSE MORFOLOOGILISEST MÄRGENDAMISEST

Reili Argus

Ülevaade

Artiklis demonstreeritakse, kuidas eesti lastekeelekorpuse ühe keeleüksuse ebajärjekindel või erinev transkribeerimine mõjutab otseselt andmete analüüsitulemusi. Põhjalikumalt on vaatluse all veakodeerimine. Igale esitatud transkribeerimisprobleemile pakutakse ka keeleandmete loetavuse ja kokkuvõtlike tulemuste adekvaatsuse printsiipe silmas pidades sobivaimaid lahendusi. Kirjeldatakse eesti lastekeele andmekogu jaoks sobiva morfoloogilise analüsaatori loomisega seotud probleeme ning peatutakse liidese loomiseks vajalikel eeltöödel, eelkõige leksikoni loomisega seonduval[1].

Võtmesõnad: lastekeel, suuline keel, transkribeerimine, kodeerimine, morfoloogiline analüsaator

[1] Tööd on toetanud ETFi grant 6151 „Koodivahetuse, eesti vahekeele ning lastekeele andmekorpuste koostamine ja üldkirjeldus" (2005–2008) ning riiklik programm „Eesti keel ja rahvuslik mälu" (2004–2008) grant R 05/01 „Koodivahetuse, vahe- ja lastekeele korpuste töötlemine ja haldamine".

Üldist: lastekeele andmete kogumise, vahetamise ja analüüsimise süsteem

Spontaanse kõne lindistamine ja transkribeerimine on ajaning töömahukas. Erinevate uurijate kogutud keelematerjali ühtedel alustel transkribeerimiseks ning töötlemiseks, võrdlemiseks ja jagamiseks on välja töötatud süsteem CHILDES (*Child Language Data Exchange System*). CHILDES-süsteemi on lastekeeleuurijad kasutanud juba üle 20 aasta, projektiga on liitunud 4500 lastekeeleuurijat, andmekogus on 130 korpust ning kasutatud materjali põhjal on ilmunud üle 1500 teadusartikli[2].

CHILDES-i andmekogus on mitut liiki korpuseid: esimese keele omandamise, teise keele omandamise, kakskeelsuse, kliiniliste probleemide uurimiseks mõeldud korpuseid. Samuti on andmekogus väga paljude eri keelte korpused. Tegemist on rahvusvahelise, mahuka ning usaldusväärse andmekoguga, mille kõige suuremaks väärtuseks võib pidada keeleandmete esitamise ühtlustatust, mis võimaldab ühelt poolt tekste kiiresti ja hõlpsalt automaatselt töödelda ning teisalt eri keete uurijail oma andmeid ja ka uurimistulemusi võrrelda.

Eesti laste suulise kõne lindistused on CHILDES-i andmekogus alates 1998. aastast. Eesti keele korpus koosneb kolmest alamkorpusest, mis on nimetatud nende koostajate järgi – Argus, Kohler ja Vija. Mahukaim neist on alamkorpus Vija, mis sisaldab 76 tunni lindistatud dialoogide transkribeeringuid. Kohleri korpus koosneb mitme lapse lasteaias lindistatud dialoogidest, Arguse korpuses on ühe lapse ja tema ema dialoogide lindistused. Kõik eesti andmekogud on ainult osaliselt, eelkõige konkreetse uurija huvidele vastavalt kodeeritud. Eesti andme-

[2] Vt ülevaadet CHILDES-süsteemist http://childes.psy.cmu.edu/intro/, 21.06.2007

kogule on plaanis lisada Tallinna Ülikoolis 2005.–2006. aastal[3] kogutud 44 tunni mahus videolindistusi, mis sisaldavad kolmikutena sündinud kolme lapse spontaanseid igapäevavestlusi oma hoidja, vanemate ja lindistajaga. Lisaks on olemas ka juba märgendatud kolmikute nn proovikorpus ehk 2000.–2001. aastal lindistatud ühe kolmikute grupi igapäevased dialoogid (5 tundi ja 20 minutit). Ka see väike andmekogu on kavas kolmikute korpusele lisada. Kolmikute kõne korpuse koostamise ja transkribeerimisega seoses ongi senisest teravamalt kerkinud esile vajadus nii praeguse eesti keele andmekogu transkribeeringu ja kodeeringu ühtlustamise kui ka keeleandmete senisest efektiivsema automaatse töötluse ja analüüsivõimaluse järele.

Hennoste (2000: 92) järgi on suulise kõne andmekogu koostamise aluseks kaks üldist printsiipi[4]: autentsus "uurjale vajalik informatsioon on säilitatud viisil, mis on tõene suhtluse loomuse suhtes" ja praktilisus, mis on põhimõte, kus "transkriptsioonis oleks välja toodud need nähtused, mida uurijal on vaja uurida ega küllastataks tekstipilti". Viimase printsiibi aluseks on omakorda ühtluse- ja järjekindluse põhimõte, mille järgi peavad üht liiki asjad olema alati ühte moodi märgendatud. Ühtne ja järjekindel märgendamine viib usaldatava tulemuseni, tulemuseni, mis oleks representatiivne ehk näitaks adekvaatselt tegelikku keelekasutust. Ühelt poolt on transkriptsioon küll juba ise alati analüüsi tulemus, tahes-tahtmata peab transkribeerija juba suulise kõne esmase ülestähendamise käigus seda kõnet analüüsima ja vastavalt esmasele analüüsile ka otsuseid tegema, kuid teisalt on transkriptsioon ka järgneva töötluse ja

[3] Kolmikute kõne korpuse loomine on seotud järgmiste programmide ja sihtfinantseeritavate teemadega: riiklik programm "Eesti keel ja rahvuslik mälu" (2004–2008) ning grant R 05/01 "Koodivahetuse, vahe- ja lastekeele admekorpuste koostamine ja üldkirjeldus" (2005–2008).

[4] Transkriptsiooniprintsiipide kohta vaata lähemalt Hennoste 2000, Hennoste jt 2004.

analüüsi põhi. Seega on usaldatavuse kõrval oluline, et keeleandmed oleksid võimalikult hõlpsasti automaatselt töödeldavad. Eelesitatud põhimõtetele toetudes on järgnevalt esitatud mõtteid, kuidas senist andmekogu ühtlustada ning mida pidada silmas uue kogu transkribeerimisel ning morfoloogilise analüsaatori ehk mor-liidese koostamisel.

CHILDES-süsteemi eesti keele andmete praegusest esitusest ja töötlemisvõimalustest

CHILDESi andmepanga dialoogide transkribeerimiseks kasutatakse CHAT-süsteemi[5] . Transkribeeritud dialoogid algavad päisega, kus antakse informatsiooni lindistuse aja, koha, osalejate, kestuse, laste vanuse, lindistussituatsiooni ning vajadusel veel muudegi asjaolude kohta (vt ka Zabrodskaja 2007: 328). Põhiridadele on paigutatud kõnelejat tähistava kolmetähelise koodi järele vestluses osalejate tegelik kõne, sõltridadele lindistaja, transkribeerija ja uurija kommentaarid või kodeeringud. Sõltridade arv sõltub eelkõige uurija täpsematest uurimiseesmärkidest. Lastekeele, nagu suulise kõne puhul üldse, ei ole sageli ainult verbaalse info põhjal võimalik aru saada, millest täpselt parasjagu räägitakse, ning seetõttu tuleb transkribeerimisel kindlasti kasutada vähemalt üht sõltrida, nimelt kommentaariridа:

NÄIDE 1[6]
*MAN: käbi, käbi .
%com: MAN leidis midagi madratsi vahelt.
(Kolmikud 2;2)

[5] Vt CHAT-süsteemi transkribeerimisjuhendit http://childes.psy.cmu.edu/manuals/chat.pdf, 21.06.2007

[6] Näited on siin ja edaspidi vormistatud nii nagu CHILDES-andmekogus, vastavalt transkribeerimissüsteemile CHAT. Iga näite alla on lisatud korpuse täpsem nimi ja vaatlusaluse lapse vanus aastates ja kuudes.

Vastavalt uurija eesmärkidele lisatakse transkribeeringusse sõltridasid; näiteks kui uurijal on kavas analüüsida kõneakte, kodeerib ta kõneaktid sõltreal %spa, kui aga morfoloogiat, siis real %mor[7].

Ka ainult põhiridadest koosnevat transriptsiooni on võimalik mingil määral oma huvidele vastavalt esitada ning automaatselt töödelda. Näiteks saab andmete paremaks jälgimiseks paigutada vaatlusaluse lapse kõnevoorud ühte, teiste kõnelejate kõnevoorud teise tulpa, koostada iga kõneleja kõnevoorude põhjal sõnasagedustabel, seda nii ühe kui ka mitme lindistuse põhjal, arvutada väljendite keskmine pikkus (MLU ehk *Mean Length of Utterance*) jne. Teiste võimaluste hulgas on näiteks ühe konkreetse muutevormi või lekseemi analüüsimiseks võimalik valida esitus, kus on välja toodud kõik selle vormi või lekseemi esinemisjuhud koos enda valitud arvu eelnevate-järgnevate kõnevoorudega. Näiteks, kasutades käsklust kwal +sbuss -w3 +w3, saab alljärgneva tulemi, kus analüüsitava lekseemi *buss* esinemisjuhud on esitatud koos 3 eelneva ja järgneva kõnevooruga:

NÄIDE 2

```
From file <c:\childes\vija\10724.cha>
----------------------------------------
*** File "c:\childes\vija\10724.cha": line 656. Keywords: buss, buss
*MOT:       pall . [+ imit]
*MOT:       istu ilusasti, vaata, mis sa õues näed .
*CHI:       bussi .
*MOT:       buss, kus buss on ?
*MOT:       mis seal sõidavad ?
*CHI:       auto .
*MOT:       auto . [+ imit]
----------------------------------------
```

[7] CHAT-süsteemi transkribeerimisreeglite kohta vt http://childes.psy.cmu.edu/manuals/chat.pdf, 2.10.2007

*** File "c:\childes\vija\10724.cha": line 662. Keyword: buss
*MOT: auto . [+ imit]
*CHI: xxx .
*MOT: vaata vaata mis seal sõidab nüüd mööda .
*CHI: buss .
*MOT: buss läheb ja näed vaata kes seal all kõnnivad .
*MOT: kes kõnnib seal ?
*CHI: onu .
(Vija 1;7)

Saadud tulemus, milles on huvi keskmes olevale vormile ja seda sisaldavale lausungile lisaks esitatud nii eelnevaid kui ka järgnevaid kõnevoore, võimaldab vormi omandamise analüüsimisel muu hulgas määrata sõnavormi välte ja arvesse võtta, ehk õigemini analüüsist välja jätta, otsesed imitatsioonid.

Transkribeerimise ebajärjekindluse mõjust analüüsitulemustele

CLAN on inglise keele põhine keeleandmete analüüsimistarkvara. Eesti keele materjali analüüsimisel tuleb seda pidevalt arvestada. Et eesti keel on suurema sünteetilisuse astmega kui inglise keel, siis ei saa näiteks sellist esmast tulemust nagu sõnasagedustabel täpsema statistika ja analüüsi alusena kuigi hästi kasutada. Sõnasagedustabelis on eesti keele puhul esitatud mitte lekseemide, vaid sõnavormide sagedus ja programm loeb näiteks lekseemi *kala* kolm vormi *kala*, *kalaga*, *kalale* eri lekseemideks (*types*). Eelnenust tulenevalt on statistiliselt vale ka väljendite keskmine pikkus (MLU) ja lekseemi-sõne indeks, mida programm automaatselt arvestab. Ka grammatiliste homonüümide ja sama kirjapildiga, kuid erinevas vältes olevate vormide eristamine tuleb programmi praegust varianti kasutades teha ära käsitsi ehk näites 2 esitatud viisil.

Kogu CHILDES-süsteemi lastekeele andmekogu on transkribeeritud kuuldeortograafiat kasutades[8]. Selline transkriptsioon ei anna küll absoluutselt autentset pilti lapse tegelikust keelekasutusest[9], ei võimalda muu hulgas (ilma kommentaarireal vastava märketa) eristada näiteks paljusid teise ja kolmanda-vältelisi vorme jne, kuid on erinevalt foneetilisest transkriptsioonist siiski mugav lugeda. Suulise kõne transkribeerimisel on oluline lähtuda loetavuse printsiibist (vt Hennoste jt 2004: 139) ja seetõttu on kuuldeortograafia kasutamine lastekeele transkribeerimisel kahtlemata ainuõige lahendus, eriti arvestades asjaolu, et vastavalt uurija vajadusele on alati võimalik transkribeeringule lisada ka täpse foneetilise transkriptsiooni rida.

Suulise kõne andmekogu automaatne analüüsimine ei ole lihtne, seda enam, et CHILDESi puhul on tegemist laste suulise keelega. On näiteks väidetud, et reeglipõhised automaatse kodeerimise süsteemid ei saagi väga hästi sellise raske keelematerjaliga, nagu seda on lastekeel, hakkama (Laakso 2005: 2). Seega esimene probleem, mis vajab läbimõeldud lahendust juba andmekogu koostamise esmasel ehk transkribeerimistasandil, tulenebki eelkõige sellest, et tegemist on suulise kõnega. Spontaanse kõne lindistused sisaldavad suulisele keelele omaseid elemente, mida ei ole vaja automaattöötluse puhul sõnedena arvesse võtta ega hiljem ka morfoloogiliselt kodeerida. Järgnevas näites on sellisteks üksusteks lapse kõnevoorus sisalduvad häälitsused *a*, *ee.*

NÄIDE 3

*EMA:	ei ja, ei ole või ?
*CHI:	ei a ee eehe .
*EMA:	jaa, jah, pane ots külge.
*CHI:	ee ee .

(Argus 1;7)

[8] Kuuldeortografia kasutamise kohta vt lähemalt Hennoste 2000.

[9] Diskussiooni transkribeerimise absoluutsest usaldusväärsusest vt Mac Whinney 2001: 3.

Eelneva näitelõigu põhjal saab arvutada lapse VKP ehk väljendite keskmise pikkuse väärtuseks 3,0, teisisõnu on lapse lausungis keskmiselt kolm sõnet. Selline tulemus ei peegelda aga kaugelt mitte adekvaatselt lapse tegelikku keelelist arengutaset, kolmest sõnast või sõnest koosnevate lausungite all mõeldakse üldjuhul ikkagi seda, et lapse lausungis esineb kolm tähenduslikku sõna. Transkribeerimisjuhend pakub probleemi lahenduseks sellist moodust: "Kui häälikujärjendit ei saa pidada lekseemiks, peaks selle alguse markeerima sümboliga &"[10]. Sümboliga & algavad häälikujäljendid jäävad VKP arvutamisel ja ka sõnasagedustabelist välja. Sellist markeerimise viisi ei ole eesti keele korpuse juures seni kasutatud, kuid uue materjali transkribeerimisel ja andmekogule lisamisel tuleks seda kindlasti teha ning kaaluda tuleks ka olemasoleva korpuse ühtlustamist.

Teine probleemiring on seotud asjaoluga, et tegemist on laste alles areneva keelekasutusega, millel on omakorda hulk erilisi tunnuseid. Näiteks sisaldab lastekeel, eriti just varases arengustaadiumis palju ühe sõnavormi kordusi:

NÄIDE 4

*EMA: nooh .
*CHI: onu, onu, onu .
(Argus 1;7)

Praeguse transkribeeringu järgi koosneb lapse lausung kolmest sõnavormist. Kui aga transkribeerida lapse lausung kaldkriipsudega, nt *CHI: onu [/] onu [/] onu [/] onu, on analüüsitulemuseks ühest sõnavormist koosnev lausung ehk teisisõnu: programm kohtleb kaldkriipsude ja sulgudega markeeritud üksuseid kui lapse püüdu üht ja sama sõnavormi välja öelda. Viimati pakutud lahendus tundub uurija seisukohalt viivat

[10] Vt http://childes.psy.cmu.edu/manuals/chat.pdf, 24.06.2007

adekvaatsema analüüsitulemuseni ning sellise transkribeerimise kasuks tasub siinkirjutaja arvamusel otsustada eelkõige perioodil, mil lapse lausungite keskmine pikkus ilma kordusteta on alles 1–2 sõnet.

Ka onomatopoeetiliste sõnade transkribeerimisel tuleb mõelda sellele, millist analüüsitulemust soovitakse saada ehk kuidas peaksid sõnasagedustabelis sellised sõnad esindatud olema. Seni ei ole onomatopoeetiliste sõnade transkribeerimisele eesti andmekogus süsteemselt lähenetud ja need sõnad kas üheks või mitmeks sõneks transkribeeritud[11] suurel määral juhuslikult. Ka teiste keelte, näiteks inglise keele andmekogude transkribeerimisel on kasutatud erinevaid variante. Kui aga onomatopoeetilisi sõnu on keelematerjalis hulganisti, ja keelelise arengu algusjärgus ehk premorfoloogilisel perioodil see nii ka on, hakkab läbimõtlemata transkriptsioon analüüsitulemusi oluliselt mõjutama. Lause keskmine pikkus paistab nende sõnade lahkukirjutamisel suurem, kui see sisuliselt on. Näiteks võib näites 5 lapse kasutatud onomatopoeetilise väljendi *kaak kaak kaak* transkribeerimisel plussmärki kasutades[12] üheks tervikuks *kaak+kaak+kaak* saada sõnasagedustabelisse mitte sõna *kaak* esinemise kolmel korral, vaid sõna *kaak+kaak+kaak* esinemise ühel korral.

NÄIDE 5

*EMA:	joonistan käo, kägu .
*CHI:	kogi .
*EMA:	ei ole kogi .
*EMA:	kägu .
*EMA:	kuidas kägu teeb ?
*CHI:	kaak kaak kaak .

[11] CHAT-süsteem ei tunnista sidekriipsu, seetõttu on transkribeerimisel kasutatud tavaortograafiast erinevat märkimise viisi, need sõnad on kirjutatud kas kokku, nt *kaakaak* või lahku, nt *kaak kaak*.

[12] Plussmärki kasutatakse CHAT-süsteemis tavaliselt liitsõnaosade kokkukirjutamisel.

*EMA: ei tee kaak kaak .
*EMA: kuku !
*CHI: kaak kaak !
(Argus 1;9)

Kindlasti tuleb aga onomatopoeetiliste sõnade kokku- või lahkukirjutamise juures arvestada ka korpuse transkribeerimisel kasutatud üldiseid pauside ja kõnevoorude üheks või mitmeks üksuseks transkribeerimise põhimõtteid. Samas tuleb transkribeerimisel peale keeleüksuste vahele jäävate pauside ehk prosoodia arvestada ka potentsiaalset analüütilist tulemit ja keeleüksuse semantilist terviklikkust. Sama tähendussisu ehk ühe lekseemi moodustavad näiteks eelmises näites kõik kolm üksust koos. Peale selle soovitab CHAT-transkribeermisjuhend panna onomatopoeetiliste sõnade lõppu sümboli @ ja selle järele tähe *o* ning lapse enda loodud keelendite lõppu sama sümboli järele tähe *c*. Selliselt markeerituna ei pea tulevikus mor-liides kodeeringureal nende sõnade vormi määrama[13].

Veakodeerimine

Täiskasvanute keelekasutusega võrreldes on lastekeele suurimaks iseärasuseks ebakorrektsete ja ebatäielike vormide rohkus:

NÄIDE 6

*CHI: ei .
*CHI: kommi ei tõi [: söönud] .
*CHI: kommi ei .
*CHI: linna [: linnas] komm .
(Argus 2;1)

Vigade transkribeerimise puhul on olulised kolm asjaolu: analüüsija peab kõigepealt nägema seda, milline oli tegelik öel-

[13] Vt http://childes.psy.cmu.edu/manuals/chat.pdf, 21.06.2007

dud sõna või sõnavorm, seejärel ka seda, milline oli see keelend, mida kõneleja soovis öelda, ning kolmandaks muudaks andmetega töö lihtsamaks, kui juba transkriptsioonis oleksid vead esimesel tasandil liigitatud. Seega peab kodeerimine sobima vigadega seotud teoreetiliselt huvitavate uurimisaspektide, nt erinevate vigade esinemisproportsioonide ja potentsiaalsete allikate analüüsimiseks.

Praegune eesti keele andmekogu on veakodeerimise osas ebajärjekindel. Kasutatud on nii põhireal n-ö tõlke lisamist (Hendriku andmekogu, osaliselt rakendatud ka Andrease andmekogus) kui ka eraldi vigade sõltrida (osaliselt kasutatud Andrease andmekogus). Transkribeerimisreeglid võimaldavad põhireal ebakorrektse vormi järele nurksulgudesse lisada vormi või sõna, mida laps kavatses öelda ehk siis täiskasvanupärase korrektse vormi. Tuleb aga meeles pidada, et CLAN-programm analüüsib sellisel juhul just kavatsetud (korrektset) vormi ehk seda vormi, mis on esitatud nurksulgudes. Seega ei saa põhireal märgendatud andmekogu põhjal koostatud sõnasagedustabelit pimesi usaldada, näiteks kui põhireal on märgitud lapse keelend *jookse* [:jookseb], on sõnasagedustabelis esindatud kuju *jookseb* ning ei ole selge, kas lapsel konkreetses dialoogis olid pöördelõpud omandatud või mitte. Probleemi vältimiseks on kasutatud erinevaid lahendusi, näiteks on leedu korpus transkribeeritud nii, et nurksulgudesse on pandud tegelik öeldud keelend ja ilma sulgudeta esitatud korrektne vorm ehk n-ö tõlge. Analüüsi lihtsustamiseks selline n-ö tagurpidi transkribeerimise võimalus küll sobiks, aga kui lisada senisele eesti keele andmekogule uus alamkorpus, mis on eelmistest erinevalt märgendatud, ajab see uurijad pigem segadusse. Teine ja mõnevõrra parem võimalus on kasutada ebatäielike vormide transkribeerimisel ümarsulge ja lisada sinna puuduv sõnaosa, näiteks *tegi musta(ks)*. Sellisel juhul saab vajadusel analüüsimisel käsklust +r2 kasutades sulgudes materjali sõnasagedustabelisse

alles jätta. Samas sobib selline märgendusviis ikkagi ainult üht liiki vigade, ärajättude märgendamiseks, näiteks asendusvigu sellisel viisil märgendada ei saaks.

Kolmas, transkribeerimisel küll kõige töömahukam, kuid kõige parem võimalus on kõik vead tähistada põhireal sümboliga [*] ja kodeerida lahti spetsiaalsel veareal %err. Põhireal järgneb sümbol [*] vigasele sõnale või sõnavormile, nt *CHI: linna [*] komm. Iga põhireal märgistatud viga tuleb kindlasti kodeerida ka veareal, sest lahtikodeerimata vigu ei oska süsteem ei liigitada ega ka analüüsida. Veareal peab olema alati sõna või sõnavorm, mis tegelikult öeldi, sümbol = ja see sõna või sõnavorm, mida kõneleja kavatses öelda. Transkribeerija võib veel lisada vea liiki märkiva koodi sümboli $ järel, vea täpse koodi ning semikooloni, mis tähistab vea kodeeringu lõppu[14].

Kõige üldisemalt võib lastekeeles esinevad morfoloogilised vead liigitada lisamiseks, asendamiseks ja ärajätuks. Järgnev näide 7 on toodud ärajätu kohta, viga on tähistatud kõigepealt morfoloogilise vea koodiga $MOR ja seejärel täpsemat liigitust näitava koodiga $LOS.

NÄIDE 7

*CHI: tegi käed musta [*].
%err: musta = mustaks $MOR $LOS
(Vija 1;8)

Ühes sõnavormis võib esineda aga korraga mitut liiki vigu. Sellisel juhul on võimalik kasutada märgendamisel mitut veakoodi. Näiteks näites 8 on esitatud sõnavorm, kus esineb nii asendamine kui ka ärajätt ning mis on tähistatud kahe koodiga, $SUB ja $LOS.

[14] Vt http://childes.psy.cmu.edu/manuals/chat.pdf, 21.06.2007

NÄIDE 8

*CHI: *tattö* [*].
%err: tättö = rattaga $MOR $SUB $LOS
(Vija 1;7)

Kui Eesti andmekogu Arguse alamkorpuses on vead veareaga täiesti varustamata, siis Andrease andmekogus on kasutatud vigade märkimiseks erinevaid mooduseid:

NÄIDE 9

*CHI: Antsu [: Andreas] läägib [: räägib] .
*CHI: kuulada [*] .
%err: kuulada=kuulata $ PHO
(Vija 2;0)

Sõnasagedustabel esitab sellisel juhul peareal n-ö tõlgitud keelendist kavatsetud ehk korrektse vormi *räägib* ning veareal kodeeritud keelendist vigase ehk tegeliku variandi *kuulada.* Morfoloogia omandamise alase uurimustöö jaoks sobib kindlasti paremini viimane variant. Vigade täpsem kodeerimine ja liigitamine võiks esialgu lähtuda Mac Whinney pakutud põhitüüpidest[15], analüüsi käigus võib neid vajadusel täpsustada.

Kõikidele eelkirjeldatud transkriptsiooniprobleemidele on võimalik leida uurija seisukohast sobivaim lahendus. Oluline on vaid see, et transkribeerimisotsused tehtaks läbimõeldult, arvestades, et kõik selle tasandi otsused mõjutavad analüüsitulemusi. Kindlasti on aga enne mor-liidese koostamist vaja ka juba olemasoleva eesti keele andmekogu veakodeerimine üle vaadata ja ühtlustada.

[15] Vt http://childes.psy.cmu.edu/manuals/clan.pdf, 21.06.2007

Morfoloogilise analüsaatori vajalikkusest, selle üldisest ülesehitusest ja koostamise erinevatest võimalustest

CLAN-süsteemil puudub senini eesti keele jaoks sobiv morfoloogiline analüsaator. Seega ei võimalda süsteem muutevormide automaatset statistikat, näiteks ei ole võimalik saada andmeid selle kohta, kui palju esineb mitmuse partitiivivorme ühes dialoogis lapse, kui palju vanema kõnes, milline on mitmuse partitiivis esinevate lekseemide sagedus kogu korpuses, ühesõnaga – distributsioonianalüüs tuleb praegu teha käsitsi. Morfoloogiliselt kodeeritud korpus annaks aga võimaluse andmeid automaatselt töödelda, arvutada mingi kindla vormi sagedusi, tuua välja näiteks kõik mingi kategooria vormid lapse vanuse järjekorras ja võrrelda üksikute vormide sagedust vastavate vormide sagedusega hoidjakeeles. Automaatse töötluse võimaldamiseks tuleb transkriptsioon varustada morfoloogilise kodeeringureaga. Näiteks näeb inglisekeelse kõnevooru morfoloogilise kodeeringuga sõltrida välja selline:

NÄIDE 10

*CHI: I want to go back.
%mor: pro|I v|want inf|to^prep|to v|go
adv|back^n|back^v|back.[16]

Kodeeringureal on iga lekseemi puhul esmalt määratud selle sõnaliigiline kuuluvus, nt verbi tähistab lühend v, noomenit n jne. Seejärel tuleb püstkriipsu järel sõna tüvi ja selle järel muutemorfoloogiat puudutav info. Kodeering on n-ö kontekstivaba, see tähendab, et esitatakse ühe sõne kõik võimalikud morfoloogilised tõlgendused. Näites 10 esinev märk ^ tähistabki

[16] Näide pärineb CLAN-süsteemi kasutusjuhendist http://childes.psy.cmu.edu/manuals/clan.pdf, 21.06.2007

kaht võrdset tõlgendamise võimalust. Õige tõlgenduse võib valida, kui kasutada spetsiaalset režiimi (*mode*). Nii tuleks näiteks morfoloogilise info sõltreal (näide 11) esitatud kõikidest võimalikest tõlgendustest valida just konkreetses lausungis realiseeruv.

NÄIDE 11

*CHI: sööb muna.
%mor: v|söö-3SG n|muna:NOM^n|muna:PARTIT^ n|muna:GEN.
(Vija 2;0)

Kogu eesti keele omandamise korpuse käsitsi morfoloogiline kodeerimine on väga aja- ja töömahukas.[17] Kodeerimissüsteem ehk mor-liides, mis baseeruks juba väljatöötatud ja kõigi keelte jaoks ühisel põhjal (*parser*), kuid oleks kohandatud siiski eesti keele morfoloogilise süsteemi järgi, genereeriks kodeeringurea aga automaatselt. Käsitsi tuleks üle kontrollida ainult mitut tõlgendusvõimalust pakkuvad kõnevoorud.

Lastekeele morfoloogia automaatse kodeerimissüsteemi töötas välja Brian Mac Whinney (MacWhinney 2000/1:104), Steven Gillis on seda täiendanud MinMOR-süsteemiga, mis sobib ka keerukama morfoloogiaga keelte tarbeks (Stephany & Bast 2007: 23). Kui transkribeerimise puhul on lastekeel juba olemuslikult keerukas materjal, siis reeglipõhise analüüsitarkvara loomise seisukohast võib lastekeel olla hoopis lihtsam kui täiskasvanute keel. Lastekeeles ja ka lastele suunatud keeles ehk hoidjakeeles ei esine kõiki keeles olemas olevaid vorme, näiteks puuduvad lastele suunatud kõnest enneminevikuvormid ja mõned ainsuse ja paljud mitmuse käändevormid (vt näiteks Argus 2005: 46). Seega võib lastekeele mor-liidese tarbeks kasu-

[17] Väite illustreerimiseks olgu siinkohal lisatud, et tunniajases lapse ja vanema vahelise spontaanse kõne lindistuses on rohkem kui 1000 kõnevooru, ainuüksi kogu Vija alamkorpuses on seega rohkem 76000 kõnevooru.

tada eesti keele morfoloogiareeglitest ainult tuumikosa. Samuti on laste- ja hoidjakeelele omane tavakeelest tunduvalt väiksem sõnavara.

Et mor-liidese keskne komponent on kõikide keelte jaoks ühine, seisab eesti mor-liidese koostaja ees põhimõtteliselt kaks suuremat ülesannet: koostada eesti keele morfoloogia kirjeldus ehk reeglifailid ning sõnastik ehk leksikonifail(id). Nii reegli- kui ka leksikonifaile võib edaspidi alati muuta ja täiendada. Morfoloogiliste reeglite kirjeldus on CLAN-programmis paigutatud kolme faili: ar.cut, cr.cut ja sf.cut. Failis ar.cut peaks olema kirjas, kuidas morfeemide kuju vaheldub või muutub ehk allomorfide kombinatoorika. Teises reeglifailis (cr.cut) on üksuste kombinatoorikat kirjeldavad morfotaktika reeglid ehk see, millised üksused millises järjestuses ja mis tingimustel võivad ühes sõnavormis koos esineda, ning kolmas fail on süntaktilise info jaoks. Failinimede ette lisatakse tavaliselt keelespetsiifiline tähis, nt saksa keele vastavad failid on saanud nimeks **ger**ar.cut, **ger**cr.cut ja **ger**lex.cut. Lekskonifail on tavaliselt jagatud eri sõnaliikide kaupa, nt adj.cut, prep.cut või n.cut. Eraldi failid koostatakse tavaliselt ka onomatopoeetiliste sõnade, pärisnimede, lapse enda leiutatud sõnade, personaal-sotsiaalsete rutiinsõnade (nt *tšauki*, *tadaa* jms) jaoks. Lisaks tüvedele paigutatakse mõnikord eraldi failidesse kõik prefiksid ja sufiksid (affix.cut) (vt Mac Whinney 2007).

Morfoloogiliste kategooriate tarbeks CLAN-süsteemi juhendis esitatud koodid sobivad suurel määral ka eesti keele jaoks. Erandiks on need kategooriad, mida eesti keeles ei ole, näiteks grammatiline sugu ja artiklite kodeerimiseks mõeldud koodid, samuti pole siinkirjutaja arvamusel vajalik kodeerida eraldi sõnaliigina modaalverbe. Kuigi on arutletud selle üle, kas markeerimata kategooriad, nagu olevik, ainsus, indikatiiv või nimetav kääne, kodeerida või mitte (nt Gavarró 2000: 6), tuleks

siinkirjutaja arvates need siiski kodeerida, kõigepealt selleks, et andmed oleksid teiste keeltega, kus seda on enamasti tehtud, võrreldavad ning ka seetõttu, et nn neutraalkategooria kodeerimata jätmisel võib tekkida näiteks grammatiliste homonüümide puhul hilisemaid analüüsiraskusi.

Leksikoni koostamisest ja MinMOR-lahendusest

Leksikonifailide koostamisel tuleb kõigepealt otsutada, mida kasutada leksikoni põhja ehk toormaterjalina. Selleks et leksikoni korpusega paremini sobitada, on kasutatud erinevaid tehnikaid (vt nt Stephany & Bast 2007: 25 või CLAN-süsteemi juhend[18]). Eesti keele leksikoni jaoks oleks üks võimalus kasutada Eesti Keele Instituudi sõnaloendeid[19]. Kõnealune sõnaloend sisaldab 100 000 sõna ja on koostatud sõnastike alusel. Kuid juba põguski pilk sõnaloendile annab alust arvata, et sellises sõnaloendis on väga palju sõnu, mida laps ja lapsega rääkiv vanem kunagi ei kasuta. Samuti on selline sõnaloend lastekeele jaoks liiga mahukas, arvestades eelkõige seda, et kõik sõnad tuleb leksikonifailis ükshaaval sõnaliigiliselt määratleda.

Teine võimalus on kasutada toormatejalina korpust ennast, teisisõnu: leksikoni võib koostada lapse ja vanema dialoogide põhjal ise. Nii sobib see kõige paremini just selle materjaliga, mida analüüsida kavatsetakse. Leksikoni võib hakata üles ehitama kahel viisil, olenevalt sellest, kas mor-analüsaatori jaoks valitakse nn põhilahendus või MinMOR-lahendus. Eri keelte puhul on mor-liidese koostamisel kasutatudki erinevaid võimalusi. Saksa keele andmekogude morfoloogilise kodeerimise tarbeks on kasutatud MinMOR-varianti, kus juba leksikoni-

[18] http://childes.psy.cmu.edu/manuals/clan.pdf, 21.06.2007
[19] Vt http://www.eki.ee/tarkvara/wordlist/, 21.06.2007

failis on kõik sõnad ka vormiliselt kodeeritud[20], inglise keele puhul on tegemist pigem analüütilisema lähenemisega ehk leksikonifailidega, kus on pelgalt tüved, ja reeglifailidega. Esimese valiku puhul võib koostada ühe keele korpuse kõikide dialoogide transkriptsioonide põhjal sõnasagedustabeli ja sealt valida esialgu näiteks sada kõige sagedasemat sõna, mis leksikoni panna. Leksikonifailid on üles ehitatud alfabeetiliselt ning sõnadele tuleb anda leksikonis ainult sõnaliigikood. Selle lähenemise järgi näeks eesti keele leksikonifail välja nii:

NÄIDE 12

koer {[scat n]}
komm {[scat n]}
kook {[scat n]}
käima {[scat v]}
loll {[scat adj]}

Inglise keele puhul ongi just sellist moodust kasutatud. Samas antakse inglise keele mor-liidese leksikonifailides ka grammatilist infot. Näiteks esitatake leksikonifailis ebareeglipäraste vormide eri tüvede täpsem kodeering:

NÄIDE 13

go {[scat v] [ir +]}
went {[scat v] [tense past]} "go&PAST"[21]

Teise valiku ehk MinMOR-analüsaatori puhul[22] tekitatakse leksikonifaili kõigepealt ainult üks kodeeritud kirje, näiteks on saksa keele leksikonifaili esimeses üksuses kodeeritud nii sõnaliik, sugu, kääne, arv:

---

[20] Samas ei ole saksa keele puhul mitte kogu grammatiline info esitatud leksikonifailis, näiteks selline nähtus nagu *umlaut* on kirjeldatud reeglifailis.

[21] Näide pärineb CLAN-süsteemi juhendist http://childes.psy.cmu.edu/manuals/clan.pdf, 21.06.2007

[22] MinMOR-lahendus sarnaneb nn eest taha analüüsiga (vt Viks 1994: 155), sest sellises variandis on morfotaktika reegilid lülitatud juba sõnastikuinfo koosseisu.

NÄIDE 14

hund {[scat N]} "hund:MASC:NOM/ACC:SG"
(Stephany & Bast 2007: 125)

Seejärel kasutatakse spetsiaalset käsklust, et tekitada ühe transkriptsiooni põhjal leksikonifaili sõnaloend, kus kõikide sõnade järel on antud vormilised lüngad, kuid kus sõnaliigi kohal on küsimärk ning puudu on ka täpsem morfoloogiline kodeering, mis tuleb kodeerijal endal lisada (vt nt Stephany & Bast 2007: 125). Mõlemal eelkirjeldatud juhul saab ja tuleb leksikonifaili hiljem täiendada.

Ühelt poolt on hea, kui leksikonifailid ei lähe liiga mahukaks, nagu nad siis, kui sisaldavad mitte tüvesid, vaid sõnavorme, võivad tahes tahtmata muutuda. Teisalt on eesti keel aga tunduvalt keerukama morfoloogiaga kui saksa keel, ka tüvevaheldusega sõnu, mille puhul tuleks nii või teisiti leksikonifailis anda vähemalt kaks tüvevarianti, on eesti keeles väga palju, samuti on palju sama fonoloogilise struktuuriga, kuid erinevatele tüvemuutusreeglitele alluvaid sõnu. Et liigselt keerukat allomorfide reeglistikku vältida, oleks ilmselt ka eesti keele tarbeks loodava mor-liidese puhul parem kasutada varianti, kus juba leksikonis sisaldub küllalt palju grammatilist infot. Kindlasti tuleb leksikonis kodeerida supletiivsed vormid. Liitsõnade ja tuletistegi esitamine leksikonis tuleb läbi mõelda: et liitsõnade moodustamist ja morfotaktikat ei saa väga hästi puhtformaalsete reeglitega kirjeldada (vt Viks 1994: 155) ja et liitsõnu on nii laste kõnes kui ka lastele suunatud kõnes vähe, on ilmselt otstarbekas esitada liitsõnad leksikonifailis.

Otsustamist vajab ka see, mis vorm valida leksikonifaili, reeglifailide koostamise seisukohast oleks parim lahendus see, kui leksikonifailis oleks esindatud tunnuseta tüvevariant, nimisõnade puhul siis ainsuse nominatiiv ja verbide puhul imperatiivis ja eitavas kõnes esinev tüvi. Täpsem piir leksikoni- ja reeglifailides antava info kohta tuleb aga fikseerida tõenäoliselt lii-

dese koostamise käigus. Kui luua esialgu väikesemahuline leksikonifail ja piiratud reeglisüsteem ning proovida, kuidas see ühe väikese dialoogilõigu peal töötab, saab ka selgemaks, milline informatsioon paigutada leksikoni ja milline reeglifaili.

Kokkuvõtteks

Eelesitatud on vaid mõned eesti lastekeelekorpuse transkribeerimise ja analüüsimisega seotud probleemid ja nende lahendusvõimalused, kindlasti tekib andmete edasise analüüsi ning mor-liidese loomise käigus neid esile veelgi. Siinse kirjutise eesmärk oli aga osutada tõsiasjale, et kõik esimese ehk transkribeerimistasandi otsused mõjutavad otseselt nii vahepealse tasandi, automaatse morfoloogilise kodeeringu kui ka viimase tasandi, uurimistulemuste kvaliteeti. Seega tasub, mõeldes neile, kes tulevikus eesti morfoloogia omandamisega tegelema hakkavad, panustada nii senise andmekogu kriitilise pilguga ülevaatamisse ja ühtlustamisse kui ka andmete automaatset töötlemist võimaldava mor-liidese loomisse.

Kirjandus

Argus,Reili 2004. Eesti keele käändesüsteemi omandamine: esimestest sõnadest miniparadigmadeni. – Emakeele Seltsi aastaraamat 2003, 49 / Toim. M. Erelt. Tallinn: Eesti Keele Sihtasutus, 23–49.

CHILDES-süsteemi üldkirjeldus. A general overview in Powerpoint format from Brian MacWhinney, http://childes.psy.cmu.edu/intro/, 21.06.2007.

CHAT-transkribeerimisjuhend, http://childes.psy.cmu.edu/manuals/chat.pdf, 21.06.2007.

CLAN-süsteemi juhend, http://childes.psy.cmu.edu/manuals/clan.pdf, 21.06.2007.

CHILDES-i keelekogud, http://childes.psy.cmu.edu/data/http://childes.psy.cmu.edu/data/, 21.06.2007.

Gavarró, Anna 2000. Proposal for a syntactic coding of Catalan for CHILDES. Report de Recerca GGT-00-4, UAB, 1–10, http://seneca.uab.es/ggt/Reports/GGT-00-4.pdf, 21.06.2007.

Hennoste, Tiit 2000. Suulise eesti keele uurimine: transkriptsioon, taust ja korpus. – Keel ja Kirjandus 2, 91–106.

Hennoste, Tiit & Koit, Mare & Strandson, Krista & Rääbis, Andriela & Valdisoo, Maret & Vutt, Evely 2004. Küsimuste ja direktiivide märgendamine eestikeelsetes infodialoogides. – Toimiv keel II. Töid rakenduslingvistika alalt / Toim. J. Lepasaar, M.-M. Sepper. Tallinna Pedagoogikaülikooli eesti filoloogia osakonna toimetised 3. Tallinn: TPÜ kirjastus, 138–155.

Laakso, Aarre 2005. On Parsing CHILDES. – Midwest Computational Linguistic Colloquium, 4/10/2005, http://cogprints.org/4204/parsing-childes.pdf, 14.10.2007.

MacWhinney, Brian 2000. The CHILDES Project: Tools for Analyzing Talk. 3rd ed. Mahwah, NJ: Lawrence Erlbaum. 2 vols.

Mac Whinney, Brian 2001. From CHILDES to TalkBank. – Research on Child Language Acquisiton / Ed. by M. Almgren, A. Barreña, M. Ezeizaberrena, I. Idiazabal, B. MacWhinney. Cascadilla: Somerville, MA, 17–34.

Stephany, Ursula & Bast, Connie 2007. Working with the childes tools: transcription, coding and analysis, http://childes.psy.cmu.edu/intro/stephany.pdf, 21.06.2007.

Zabrodskaja, Anastassia 2007. Vene-eesti koodivahetuse korpus: kodeerimispõhimõtete väljatöötamine. – Eesti Rakenduslingvistika Ühingu aastaraamat / Toim. H. Metslang, M. Langemets, M.-M. Sepper. Tallinn: Eesti Keele Sihtasutus, 321–338.

Viks, Ülle 1994. Eesti keele morfoloogiline analüsaator. – Keel ja Kirjandus 3, 150–163.

Morphological coding of Estonian child language database

Summary

The focus of the paper lies on different problems connected to transcription and analysis of the Estonian CHILDES database. The paper illustrates how the influence of inconsistent transcription and coding of different elements like onomatopoetic words, repetitions and errors can lead to inadequate results in a research. The main emphasis of the paper is on the coding of morphological errors. The most suitable and suggested solution to error coding would be the adding of special dependent tier *err%* containing the morphemic translation of the incorrect word form. The existence of the dependent tier is the base for further automatic error analysis with CLAN-programs. The present paper introduces some suggestions and ideas on the creation of the morphological analyzer for the Estonian CHILDES database.

The MOR-program will automatically generate a *%mor* tier for each main tier. In order to create the program one needs to configure both grammar files and lexicon files. Estonian language has a very complicated morphological system with a large set of different stem variants and allomorphs. It is important to identify, to which extent the grammatical information should be stored in the lexicon files and to which extent in the rule files. The article finds that suppletive forms should be a part of the lexicon file, but the information regarding different stem variants remains the question of further discussion.