Tallinna Ülikool
Informaatika Instituut

Kõnesalvestuste transkribeerimine laste kõne korpuse näitel

Transcription of Speech on the Base of Children's Speech Corpus

Bakalaureusetöö

Autor: Ethel Maarits
Juhendaja: Erika Matsak

Autor: ..... ,, 2011
Juhendaja: ..... ,, ... ,, 2011
Instituudi direktor: ..... ,, ... ,, 2011

Tallinn 2011

Autorideklaratsioon

Deklareerin, et käesolev bakalaureusetöö on minu töö tulemus ja seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud.

....

(kuupäev) (allkiri)

Sisukord

Sissejuhatus

Kirjakeel uueneb pidevalt iga päevaga ning seepärast on käesolev töö aktuaalne. Käesoleva töö eesmärk on luua laste kõne korpus, mis vastaks Childes andmebaasi ülesehitusele ning mis oleks arvuti tavakasutajale lihtsasti kasutatav ja arusaadav.

Töö sihtrühmaks on erineva valdkonna teadlased (kasvatusteadlased, keeleteadlased, psühholoogid jms). Töö tulemusel loodav korpus aitab teadlastel uurida laste kõne ning annab adekvaatsema ülevaate kõnest. Autori töö kuulub keeletehnoloogia valdkonda.

Käesolevas töös kasutab autor arendusuuringut.
Käesolev bakalaureusetöö on jaotatud nelja peatükki, millest esimene tutvustab lühidalt korpust ja korpuseid Eestis. Teises peatükis tutvustab töö autor Childes andmebaasi ja CLAN programmi ja nende kasutusvõimalusi. Kolmandas peatükis tutvustab autor WordPress´i ja võimalusi korpuse loomiseks. Neljandas peatükis kirjeldab autor loodavat korpust, toob välja automaatse transkribeerimise võimalused ja kõnesalvestuste transkribeerimise kirjeldus. Samuti on selles peatükis välja toodud korpuse loomine ja parendamine.

Eesmärgi saavutamiseks transkribeerib autor laste kõne helifailid, sisestab transkribeeritud failid CLAN programmi ning loob WordPress´i kasutades laste kõne korpuse.

Autor kasutab oma töös teemakohast informatsiooni kogumist ja analüüsi vastavalt püstitatud teema vajadustele.

1. Korpus

Korpus on arvutisse viidud tekstide kogum, mis on valitud kindlate kriteeriumite alusel ja esineb ühtses elektroonilises formaadis. (Muischnek, Orav, Kaalep, Õim, 2003, 9)

Korpuseid kasutatakse leksikograafias, lingvistikas, arvutilingvistikas ning keeletehnoloogias (Stubbs, 1996). Seejuures on väljatöötatud palju erinevat tarkvara, mis võimaldab analüüsida nii teksti (Scott 1997) kui ka kõne, samuti kõne alusel uuritakse isegi emotsioone (Altrov, 2008), andes uurijale tagasisidet statistiliste omaduste kohta. Nii näiteks teksti puhul on võimalik leida sagedasemad sõnad ja sõnakooslused, uurida erinevate žanride omapära jne. Kõne puhul on võimalik uurida hääldamise sagedasi mustreid ning kasutada neid kõnetuvastamisel (Alumäe, 2007), kõnesünteesil (Mihkla, 2007) ning samuti aktsendi uuringutel (Meister, 2009).

Esimesed korpused loodi 1960ndatel aastatel. Nendeks olid Browni korpus USA-s ja Lancaster-Oslo/Bergeni korpus Inglismaal. Aastakümnetega on korpuste mahud ja kasutusvõimalused pidevalt suurenenud. (Muischnek, Orav, Kaalep, Õim, 2003)

Suuremate korpuste sõnade maht ulatub miljardite sõnadeni. Maailma kõige suurem keelekorpus, Korpora geschriebener Gegenwartssprache, asub Saksamaal. 29.03.2011 seisuga on korpuses 4,1 miljardit sõna. (allikas: http://www.ids-mannheim.de/kl/projekte/korpora/)

1.1. Millised korpused on Eestis olemas?

Eesti keelekorpuste kohta on loodud internetilehekülg (Eesti Keele Keeletehnoloogiline Tugi), kust on võimalik leida infot käimasolevate ja lõppenud projektide kohta: http://www.keeletehnoloogia.ee/projektid
Kõik need korpused on vabalt kättesaadavad mittetulunduslikel eesmärkidel. Eesti kirjakeele korpused ei kajasta luulet ega draamat.

Korpuste eesmärgid on väga laiahaardelised - masintõlke kvaliteedi parandamine, kõnes kajastuvate emotsioonide uurimine ja taristu loomine. Need on ainult vähesed teemad, mida korpused hõlmavad ning kõik see on kättesaadav kõigile eesti keeletehnoloogia huvilistele. (allikas: http://www.keeletehnoloogia.ee/projektid)

Korpustest on kasu ka keele õppimise puhul. Eesti keele kui võõrkeele õppe arengu märkimisväärseks teguriks on muutunud keelekorpused. Infotehnoloogia vahendite arengu tõttu on võimalik võõrkeele korpust paremini uurida ja katsetada. (Kitsnik, 2006)

2. Childes andmebaas

CHILDES (Child Language Data Exchange System) on laste keele andmebaas, mille eesmärgiks on talletada ja uurida laste keele omandamist erinevate maailma keelte näitel. Andmebaas koosneb erinevate inimeste poolt lindistatud suhtlussituatsioonide transkriptsioonidest ja hõlmab väga mitmete keelte näiteid.
Childes võimaldab uurida kõnekeele vastastikust toimet.

Childes andmebaas asutati 1984.aastal Carnegie Mellon Ülikooli psühholoogia osakonnas. Loomise hetkel koosnes Childes tiim selle loojast Brian MacWhinney´st ja kahest programmeerijast – Leonid Spektor ja Franklin Chen.

Tänaseks on Childes andmebaasil rohkem kui 4500 liiget ja üle 130 korpuse. Need numbrid kasvavad iga päevaga. Childes´i kohta on avaldatud rohkem kui 1500 artiklit.

Childes andmebaas on vajalik selleks, et seal leiduvaid faile kuulates ning vastavate transkribeeritud tekste analüüsides saavad teadlased järeldada väga erinevaid aspekte. Lalisemise ühtne muster, varajaste sõnade universaalne järjestus, ühtlane sõnavara spurt, laste individuaalsed rääkimise stiilid, laste reageeringud vanemate ümbersõnastatud fraasidele - need on ainult vähesed valdkonnad millele antud andmebaasi abil üritatakse vastuseid leida.
Childes andmebaas asub internetiaadressil: http://childes.psy.cmu.edu/
(allikas: http://childes.psy.cmu.edu/intro/utalam.ppt)

2.1. CLAN

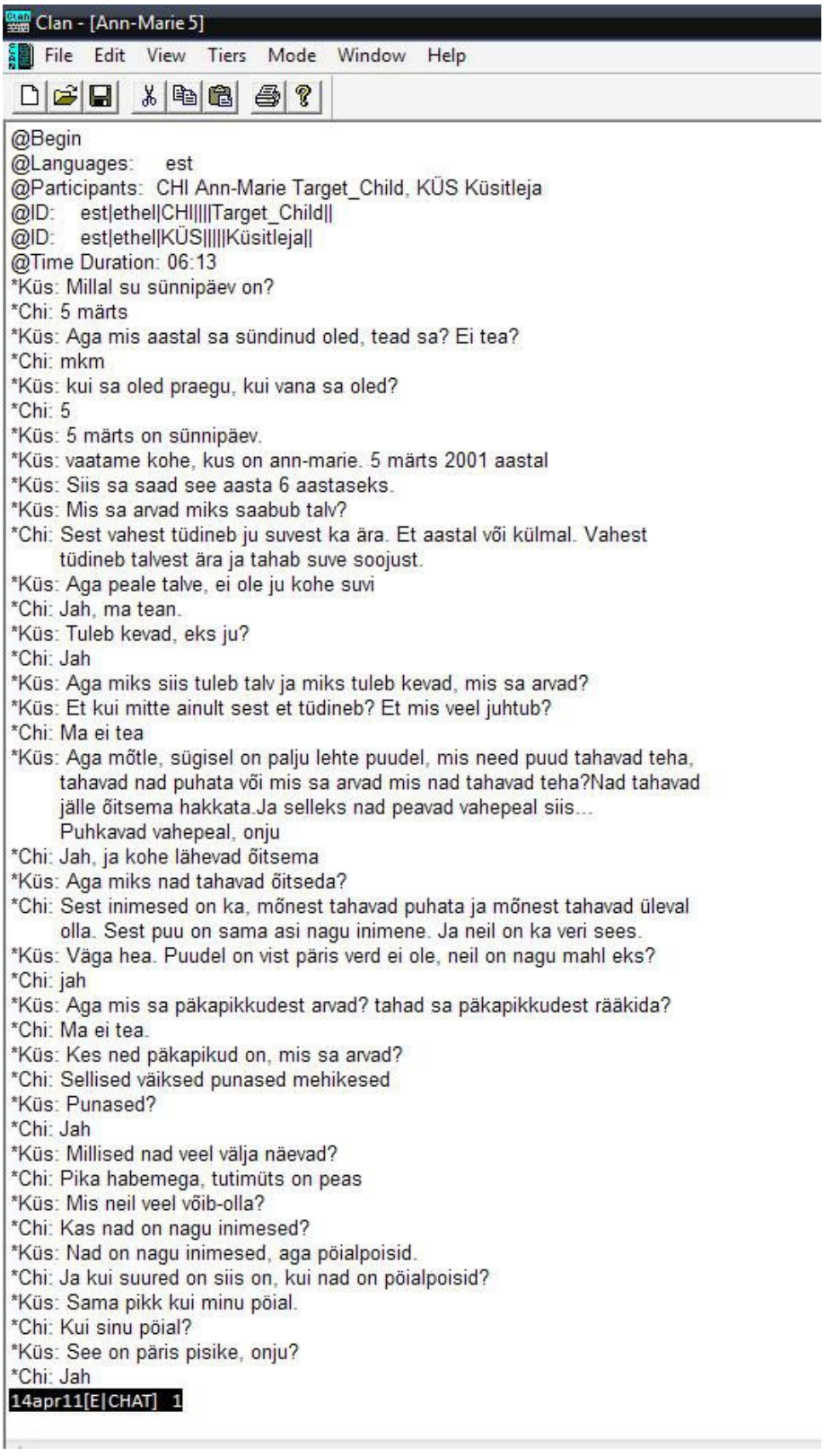
CLAN (Computerized Language Analysis) on programm, mis on loodud spetsiaalselt Childes andmebaasi andmefailide analüüsiks. CLAN on Childes andmebaasi standardne formaat.

CLAN pakub palju võimalusi lubades kasutajal kuulata helifaile ning lugeda samal ajal teksti. Lisaks loendada sõnu ja morfeemide suhtarvu ütluste suhtes, otsida andmeid määratletud sõnade kombinatsioonidele, märgijadadele või spetsiifilistele

sõnadele. CLAN failide vaatlemiseks oma arvutis tuleb alla laadida vabavaraline tasuta tarkvara http://childes.psy.cmu.edu/clan/

(allikas: http://childes.psy.cmu.edu/intro/utalam.ppt)

Näide CLAN tekstifailist (.cha) (Pilt 1).



Pilt 1 CLAN

3. WordPress

WordPress on vabavaraline avatud lähtekoodiga blogi haldamise süsteem, mis loodi 2003.aastal ühe bitilise koodina, et parandada igapäevast tüpograafiat. WordPress sündis soovist luua elegantne, hästi disainitud personaalne kirjastamise süsteem, mis töötaks PHP skriptimiskeele ja MySQL andmebaasi peal. Tänaseks on WordPress´i lehte alla laetud rohkem kui 7 miljonit korda.

Koos paljude lisade, teemade ning moodulitega on WordPress kohandatav erinevate vajaduste jaoks. Esialgu oli WordPress mõeldud blogide haldamise süsteemina, kuid hiljem arenes välja täielikuks sisuhaldussüsteemiks. Hetkel on viimaseks ametlikuks versiooniks 3.1.2, mis avalikustati 26.04.2011. (allikas: http://wordpress.org/about/)

WordPress´i saab hõlpsalt allalaadida ja paigaldada ning mõningate tehniliste teadmiste ja oskustega on soovitud lehtede loomine kiire ja sujuv.
WordPress´i paigaldamiseks on vaja:

- ligipääsu kasutaja veebiserverisse
- tekstiredaktorit
- FTP klienti
- veebibrauserit

(allikas: http://codex.wordpress.org/Installing_WordPress)

3.1. Ülevaade WordPress´i võimalustest korpuse loomiseks

WordPress´il on palju võimalusi internetilehekülje loomiseks.
Autor pidas antud korpuse loomisel kõige mõistlikumaks kasutada „Pages" (lehed) lahendust. Sellise lahendusega on korpuse alajaotused väga selgelt eristatud ning lihtsalt ülesleitavad.

Iga külastaja, kes soovib lisada oma faile korpusele, peab kõigepealt ennast registreerima kasutajaks antud lehel.
WordPress´is saab ära määrata, millised õigused uuele kasutajale vaikimisi antakse. Valida on „editor", „author", „contributor" ja „subscriber" vahel.

„Editor” saab avaldada ja hallata oma ning ka teiste postitusi ja lehti.

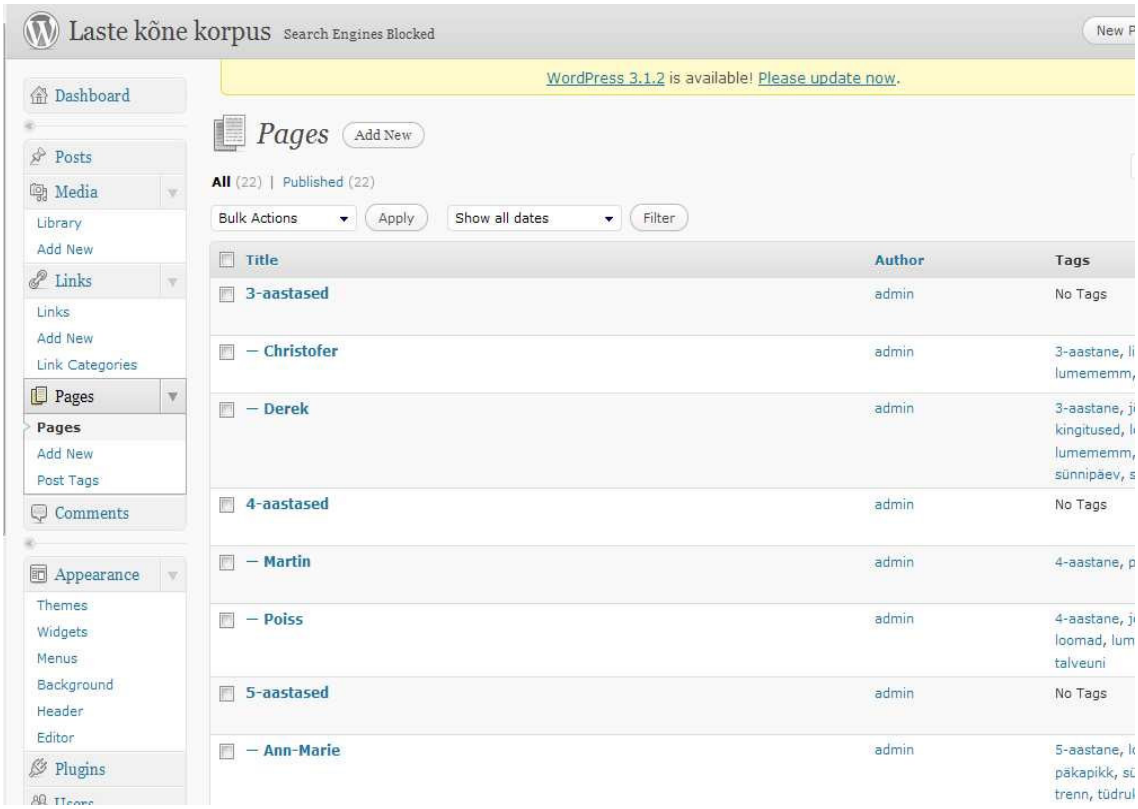
„Author” saab avaldada ja hallata ainult enda postitusi.

„Contributor” saab kirjutada ja hallata oma postitusi, kuid ta ei saa neid avaldada.

„Subscriber” saab ainult oma profiilis muudatusi teha.

Postitused („Posts”) ilmuvad esilehele üksteise alla, vastavalt avaldamise kuupäevale. Kõige uuemad kõige esimestena. See on tüüpiline lahendus blogidele.

Lehed ehk „pages” on kogu aeg staatiliselt paigas ja neile saab lisada alamlehti.



Pilt 2 WordPress´i admini vaade

4. Loodav korpus

Tallinna Ülikooli Informaatika Instituut varustas autorit tehnilise lahendusega, kuhu oli võimalik luua loodav korpus mis jääks ka edaspidi kõigile kättesaadavaks. Laste kõne korpus asub internetiaadressil: http://minitorn.cs.tlu.ee/korpus/

Loodav korpus peaks aitama keeleteadlastel, kasvatusteadlastel, psühholoogidel soovi korral uurida erinevas vanuses laste kõnekeelt.

Juhendaja varustas autorit doktoritöö käigus salvestatud laste kõnesalvestustega. Loodav korpus pidi vastama Childes andmebaasi ülesehitusele, olema lihtne ja arusaadav ka arvuti tavakasutajale.

4.1. Automaatse kõnesalvestuse transkribeerimise võimalused

Autori tööd oleks lihtsustanud kõvasti automaatne transkribeerimine. Automaatne transkribeerimine on veebibrauseris tehtud võimalikuks järgneval lehel: http://www.phon.ioc.ee/dokuwiki/doku.php?id=projects:tuvastus:veebituvastus.et Kasutaja laeb helifaili üles serverisse ning valib, kas soovib sellest loodavat tekstifaili saada veebist või meilile saadetuna.

Autor proovis mitme helifailiga automaatset transkribeerimist, kuid loodetud tulemust ei saanud. Helifail peab olema täiesti müravaba ning lapse jutt peaks olema räägitud otse mikrofoni. Need tingimused on aga lasteaias rahmeldavale lapsele peaaegu võimatud. Seega autor automaatset transkribeerimist kasutada ei saanud.

Üldiselt peaks helifailile vastav tekstifail tulema koheselt, aga kui helifailis juhtub olema liiga palju müra, siis veebileht tulemust ei anna. Seda kinnitas ka lehe omanik, kes lubas edaspidi arvestada sellega, et vea puhul kasutajatele vastav teade kuvataks.

4.2. Kõnesalvestuste transkribeerimine

Helifailid on erineva pikkusega, 1 kuni 60 minutini. Salvestistes on üks kuni mitu last. Osades failides on väga hästi kuulda laste teksti, kuid mõnel on väga suur taustamüra, nt lasteaia rühma teiste laste mänguhoos rääkimine.

Autor kasutas oma töös manuaalset transkribeerimist (ümberkirjutamist).

Kui helifail transkribeeritud, tuli transkribeeritud tekst sisestada CLAN faili.

4.3. Korpuse loomine

Autor kategoriseeris helifailid laste vanuste järgi gruppidesse (Pilt 3). Gruppideks on: 3-aastased, 4-aastased, 5-aastased, 6-aastased ja lastevahelised vestlused. Viimases grupis asuvad helifailid, kus vestlevad kaks või rohkem lapsi, kelle vanus pole teada.

Pilt 3 Korpuse grupid

Grupi nime peale liikudes avaneb rippmenüü, kust saab lehe kasutaja valida millist faili vaadelda soovib. (Pilt 4)

Pilt 4 Gruppide alamjaotus

Igal lehel on üleval ja all lingid heli- ja CLAN-failile (Pilt 5).

Pilt 5 Heli-ja tekstifaili lingid lehel

Lehe külastajal/kasutajal on võimalik helifaile kuulata ja/või alla laadida. Helifaili on võimalik salvestada kasutades hiire paremklahvi ning valides „Save Link as".
Lisaks on võimalik kasutajal lugeda ja/või alla laadida CLAN faile, mis on pakitud zip-formaati.

Iga külastaja, kellel on heli- või tekstifaile ning soovib neid ka teistega jagada, saab seda teha, registreerides end lehel kasutajaks (Pilt 6).

Esialgu saab iga liitunud kasutaja teha postitusi ning lehe administraator tõstab need õigetesse kategooriatesse.

Tulevikus on plaanis, et iga kasutaja saab oma postitusi ja lehti ise hallata.

Pilt 6 Külastaja registreerimine kasutajaks

Sildid

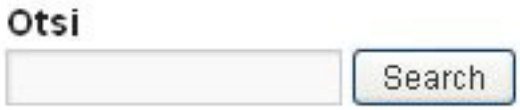
Iga tekstifaili lõpus on märgistused erinevate siltidega (Pilt 7). Sildid on lahterdatud laste vanuse, laste soo ning teemade järgi, millest lapsed helifailis räägivad.

Kõik sildid kuvatakse staatiliselt igal korpuse lehel ning on omakorda varustatud numbritega, mis näitavad kui palju ühte silti kasutatud on. Selle põhjal saab teha statistikat, millest lapsed kõige rohkem rääkisid.

Pilt 7 Sildid

Otsing

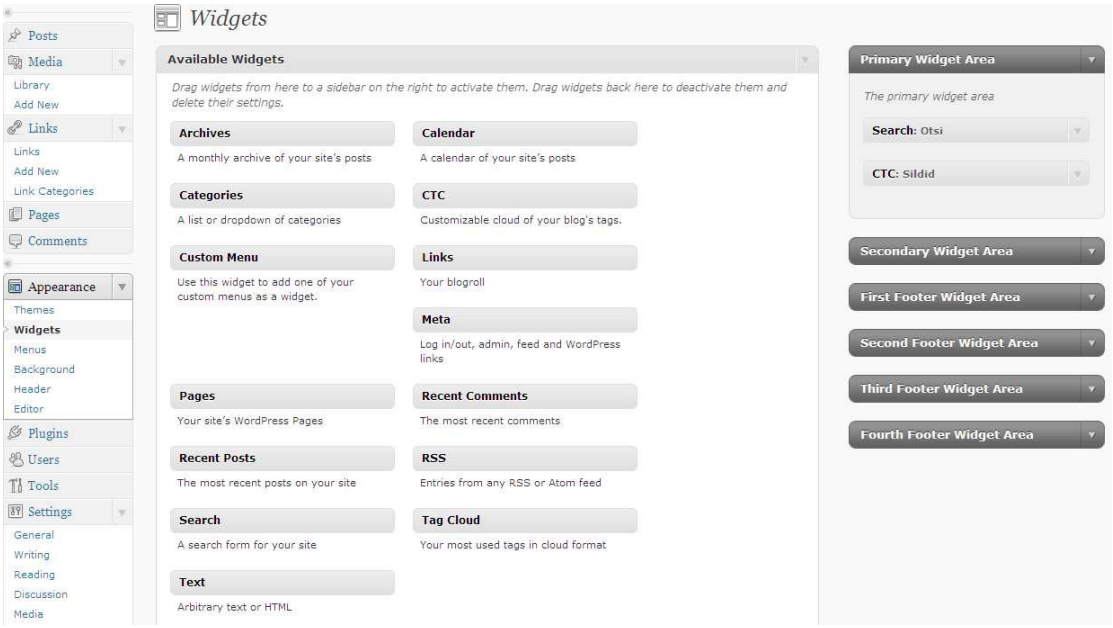
Korpus on varustatud otsimise funktsiooniga (Pilt 8). Kasutaja saab otsida märksõna järgi, kas lapsed on rääkinud otsitaval teemal.



Pilt 8 Otsi funktsioon

4.3.1. Korpuse disaini teostus

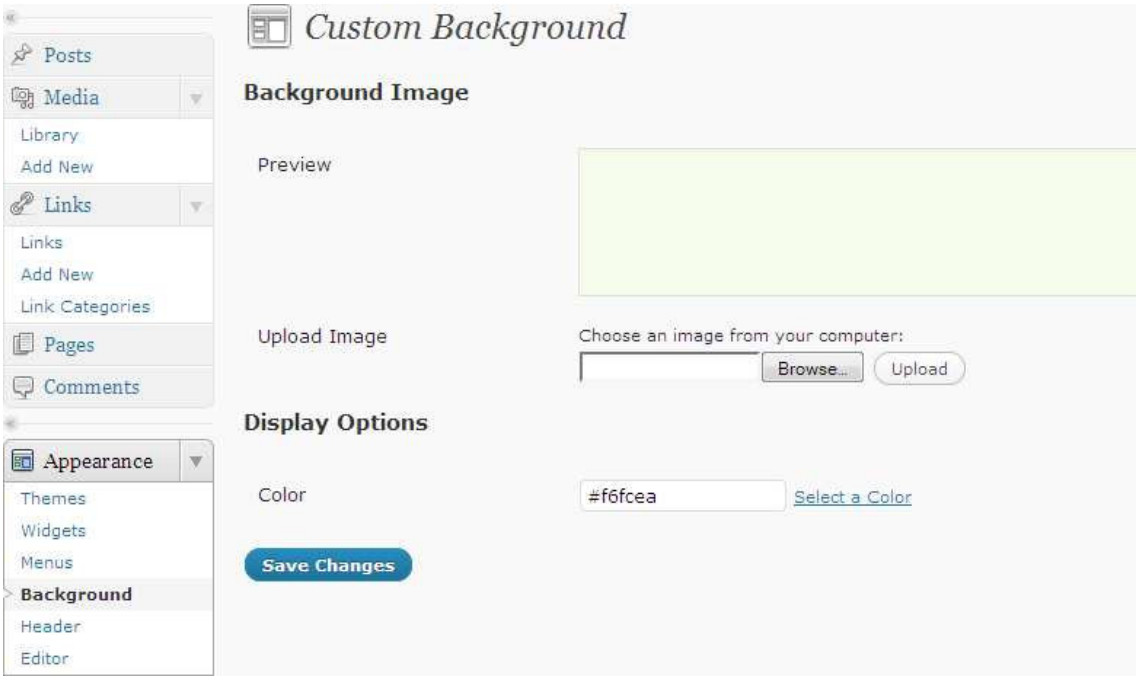
WordPress on varustatud vaikimisi mitmete moodulitega (Pilt 9), mida kasutaja soovi korral saab kas lisada või ära võtta. Lehel saab kuvada näiteks kasutajate poolt tehtud viimaseid kommentaare, postituste kategooriaid, kalendrit, arhiivi, uudistevoogu ja palju muud.



Pilt 9 Moodulid

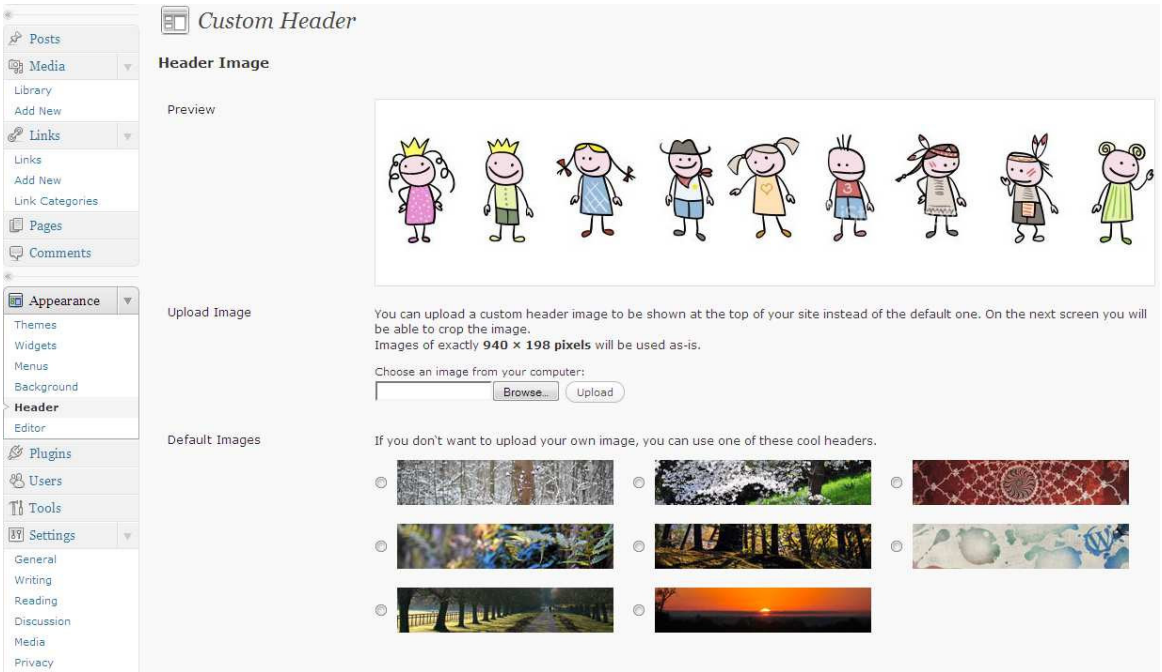
Töö autor leidis, et enamikke vaikimisi pakutavatest moodulitest pole antud korpuse puhul vaja ning eemaldas need. Moodulitest jäi otsingu ala ning plugin´atest sildistamise tööriist. Sildistamise plugin´a lehele installimiseks pidi autor ühendust võtma lehe haldajaga, kes ka autorit lehega varustas.

WordPress´i välimuse („Appearance”) alt saab valida, kas soovitakse lehe taustaks (Pilt 10) panna pilt või valida selleks värv. Autor valis selleks neutraalse helerohelise värvi.



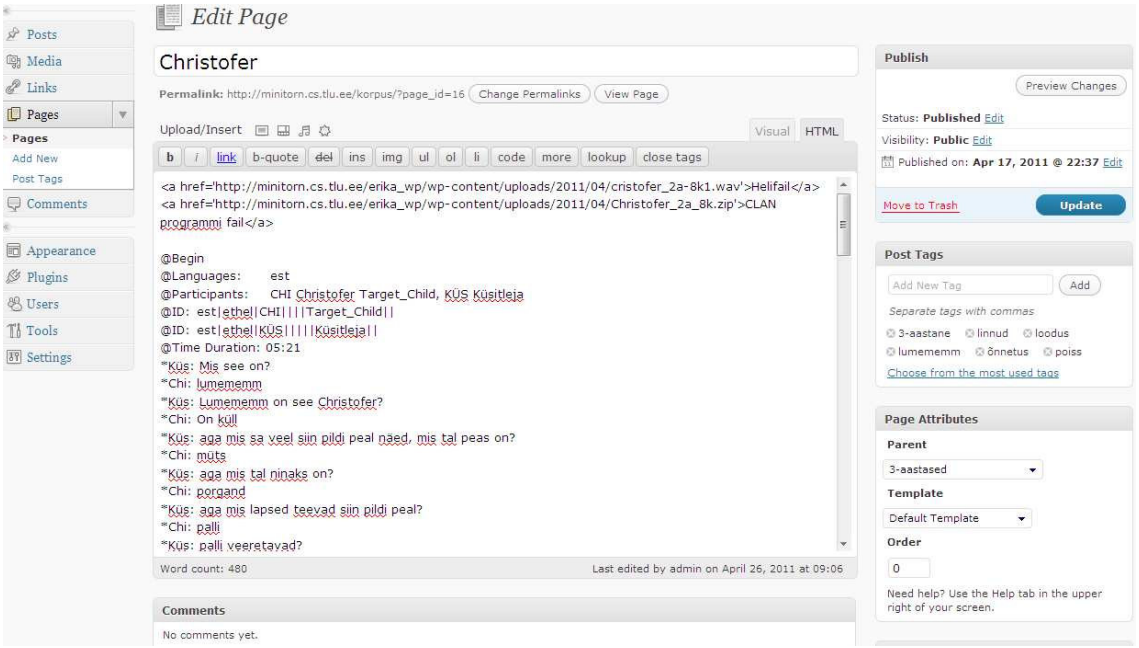
Pilt 10 WordPress´i lehe tausta valik

Valiku Päis („Header”) alt saab lehe koostaja valida pildi lehe päiseks. Vaikimisi on pakutud ka paar pilti WordPress´i poolt. Kuna töö autor soovis, et korpuse kujundus oleks omanäoline, siis lõikas ja kleepis autor lehe päiseks kokku uue pildi (Pilt 11).



Pilt 11 Päis

Uue lehe lisamiseks tuleb valida Pages -> Add New. Lehed („Pages") all asuvad kõik korpuse sisulehed (Pilt 12). Faili pealkirjaks pani korpuse autor lapse nime. Sisukasti läks tekstifail ning lingid nii heli- kui ka tekstifailile. „Upload/Insert" alt saab sisestada pilte, videosid, helifaile kui ka muud meediafaile. Autor kasutas helifailide lisamiseks „Add Audio"t ning zip-formaadis failide üleslaadimiseks „Add Media" ikooni.

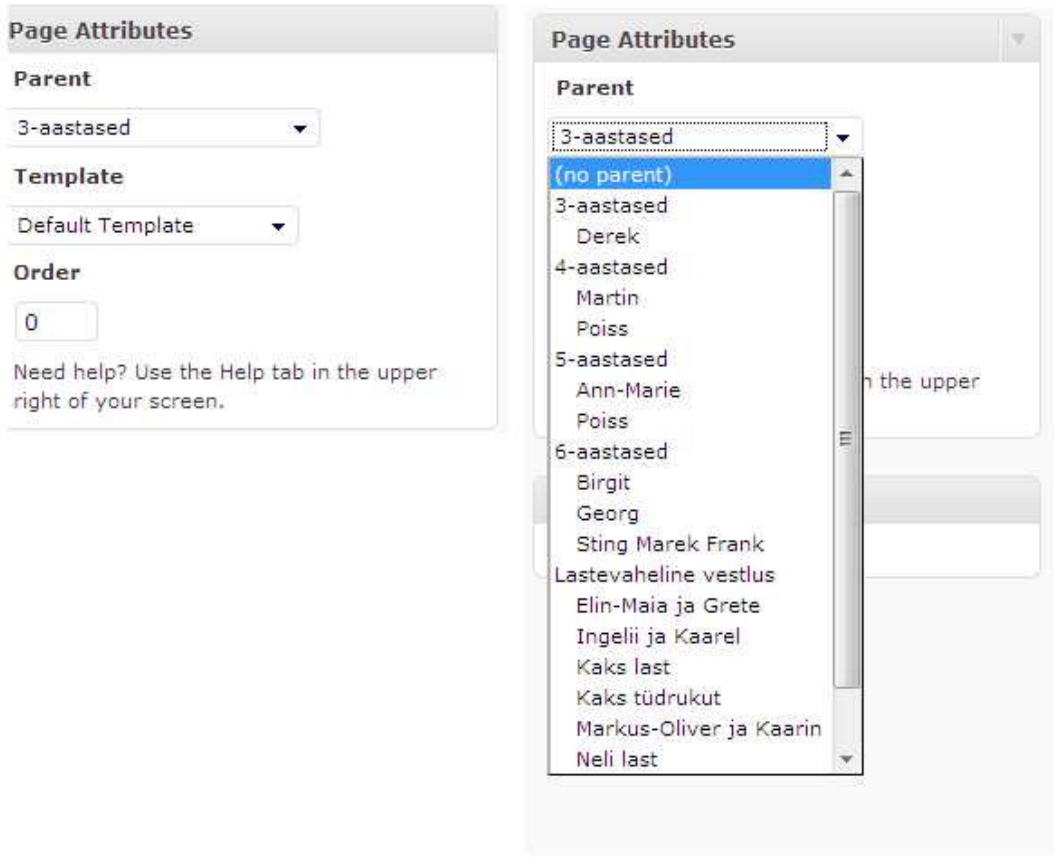


Pilt 12 Sisuleht

Igale lehele saab lisada silte (Pilt 13). Saab lisada varem mitte esinenud silte. Silte saab lisada ükshaaval kui ka mitu korraga. Korraga lisades tuleb sildid üksteisest komadega eraldada. Samuti saab ka valida varem esinenud siltide seast. Kõik sildid kajastuvad siltidepilves (Pilt 7).

Pilt 13 Siltide lisamine

Koostatud laste kõne korpusel on mitmetasandiline menüü. Lehe lisamiseks ülemlehe alla, tuleb valida lehele vastav „Parent"(Pilt 14). Järjestus („Order") määrab ära lehtede järjestuse menüüs. Kui kaks lehte on sama numbriga, siis süsteem järjestab lehed tähestiku järjekorras.



Pilt 14 Menüü ja alammenüü valimine

4.4. Korpuse testimine

Valminud korpus anti testimiseks Tallinna Ülikooli filoloogidele, et nad annaksid oma eksperthinnangu.

Eesti keele võõrkeelena professori Pille Esloni, eesti keele kui teise keele ja keelepoliitika osakonnast, sõnul meeldis talle korpuse disain. See oli esimene asi, mis talle kui kasutajale silma hakkas. Teiseks positiivseks aspektiks tõi ta välja korpuse selge jaotumise tunnuse alusel. Positiivsena märkis veel, et korpus on vabalt kasutatav ja arendatav. Veel tõi positiivsena välja selle, et helifail ja tekstifail jooksevad paralleelselt. Pille Esloni sõnul vajab veel arendamist failide glossimine ja suulisele kõnele omane märgendamine, mis peaks olema tulevikus tehtav.

Korpust testis ka Tallinna Ülikooli Informaatika Instituudi süsteemiadministraator Tanel Toova. Positiivsena tõi Toova välja selle, et kasutatakse WordPress´i blogimootorit, mis on ülesande lahenduseks piisav ning adminstreerimise seisukohast turvalisem ja lihtsamini hallatav kui iseseisvalt kirjutatud veebikeskkond. Toova sõnul on autor valitud blogimootorit kasutanud küllaltki oskuslikult – muudetud on kujundust, eemaldatud mittevajalikud elemendid ning lisatud või modifitseeritud vajalikke. Üldmulje kujundusest ja kasutajaliidesest on üldiselt hea – kasutajaliides on lihtne ning vajalik info on kiiresti leitav.
Toova toob ka välja soovituse, mida võiks muuta lehe juures. Lehtedele võiks lisada lingid lehe jaotise all olevatele postitustele. Menüüst on need kättesaadavad, kuid kui kasutaja klikib menüüs lehe elemendile, siis saab kasutaja teada, et “Siit võid leida [number] aastaste laste kõnesalvestused”.
Veel toob Toova välja selle, et uued registreeritud kasutajad saavad koheselt postitada ükskõik mida suvalise ajajaotuse alla. Selline võimalus aga tähendab veebikeskkonna kohest risustumist kõikvõimaliku spämmiga. Toova soovitab olukorra lahenduseks: uued kasutajad luuakse vaikimisi “contributor”, mitte “author” privileegiga. “Contributor” juhul saavad uued kasutajad küll postitusi kirjutada, aga mitte avaldada. Postitused peaks soovi korral avaldama lehe administraator. Tulevikus saab usaldusväärsete kasutajate õigusi käsitsi laiendada.

Samuti testis korpust Tallinna Ülikooli esimese aasta lingvistikadoktorant Helen Kõrgesaar. Kõrgesaare sõnul on lehe ülesehitus väga hea, ilus ja selge, eriti meeldis talle korpuse disain. Lehe parendamiseks tõi välja soovituse, et lehe külastaja võiks saada tekstifaili lugemisel samaaegselt helifaili kuulata.

Kõrgesaarele meeldis, et lehe juures kõik töötab, tekstifailid avanevad ning helifailide avamine pole ka keeruline.

Miinusena tõi Kõrgesaar välja selle, et helifailidest on välja kirjutatud ainult tekstiliselt korrektsed laused, laste komistamisi kõnes pole kirja pandud. Kui laps ütleb „vares" asemel „vaes", siis peakski nii failis kirjas olema. Kõrgesaar arvab, et keeleliselt korrektsetena ei anna failid adekvaatset ülevaadet korpusest. Kõrgesaar pakub, et failid tuleks uuesti litereerida.

4.5 Korpuse parendamine

Autor võttis arvesse testinute ettepanekuid ning võimaluse korral viis korpuse ülesehitusse muudatused sisse.

Pille Esloni tehtud ettepanekut suulist keelt märgendada polnud võimalik ajanappuse tõttu teha. Kuid on plaanis tulevikus leida võimalused, kuidas seda teha saaks.

Tanel Toova soovituse ära vahetada kasutaja vaikimisi privileeg “authorist” “contributor”iks viis autor koheselt sisse, kui soovitus tuli.
Toova teist soovitust, lisada lehtede vanuse jaotisele laste failide lingid, et kasutaja saaks ka vanusegrupi esilehelt soovitud lindistused kätte, võttis autor arvesse ning viib pakutud muudatused sisse korpusesse.

Helen Kõrgesaare soovitust laste failid uuesti litereerida võetakse arvesse ning muudatused viiakse lähiajal sisse.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärgiks oli luua laste kõne korpus, mis oleks külastajatele lihtsasti kasutatav ja arusaadav. Korpus pidi olema ülesehituselt sarnane Childes andmebaasile. See eesmärk sai täidetud.

Antud töös autor tutvustas lühidalt korpust ning korpuseid Eestis.

Samuti tõi autor välja lühitutvustuse Childes andmebaasist ning andmebaasiga kasutatav CLAN programmist.

Autor tõi oma töös välja tutvustuse WordPress´i võimalustest korpuse loomiseks.

Valminud korpus anti testimiseks Tallinna Ülikooli filoloogidele.

Korpust testinud inimesed leidsid, et korpus on lihtsasti kasutatav, hästi ülesehitatud. Toodi välja korpuse kujundus, mis kõigile meeldis. Testijad tegid ettepanekuid, kuidas korpust parendada saaks. Vastavalt nende ettepanekutele viis autor võimalusel parandused sisse.

Korpust saavad edaspidi kasutada kõik, kellel selleks soov on. Korpuse internetileheküljel saab külastaja ennast kasutajaks registreerida. Ning kui kasutajal on helifaile või tekstifaile, siis saab neid korpuse lehele lisada.

Keeleteadlased saavad korpuse põhjal uurida laste kõne.

Summary

The goals of this bachelor thesis are to give an overview about the corpus and to create a childrens speech corpus using WordPress, which would be easy to use on an everyday basis.

To accomplish the goal and create a children's speech corpus, the author transcribed children's speech files and inserted them into the CLAN program.

In the first chapter the author gives a short overview about corpuses. In the second chapter the author presents the Childes database and the CLAN program. In the third chapter WordPress and it´s possibilities to create a corpus are described . In the fourth chapter the author presents the children´s speech corpus, which was made for this bachelor thesis.

The corpus was given to a philologist for testing. They thought that the corpus was well built and easy to use. The design of the corpus was also given a good evaluation.

The linguists can use the children´s speech corpus in order to research children´s speech.

Kasutatud kirjandus

About WordPress. Loetud Internetis 15.aprillil 2011 aadressil http://wordpress.org/about/

Altrov, R. (2008). Eesti emotsionaalse kõne korpus: teoreetilised toetuspunktid. *Keele ja kirjandus* 4, 261-271.

Alumäe, T. Automatic compound word reconstruction for speech recognition of compounding languages. Proceedings of NODALIDA 2007.

Childes system overview. Loetud Internetis 15.aprillil aadressil http://childes.psy.cmu.edu/intro/utalam.ppt

Installing WordPress. Loetud Internetis 15.aprillil 2011 aadressil http://codex.wordpress.org/Installing_WordPress

Kitsnik, M. (2006). Keelekorpused ja võõrkeeleõpe. Eesti rakenduslingvistika aastaraamat

Korpora geschriebener Gegenwartssprache. Loetud Internetis 8.aprillil 2011 aadressil http://www.ids-mannheim.de/kl/projekte/korpora/

Käimasolevad EKKTT projektid. Loetud Internetis 12.märtsil 2011 aadressil http://www.keeletehnoloogia.ee/projektid

Meister, L. (2009). Eesti vokaalikategooriate piirid vene ja eesti emakeelega kõnelejate tajuruumis. In: Eesti Rakenduslingvistika Ühingu aastaraamat 5 = Estonian Papers in Applied Linguistics 5: (Toim.) Metslang, Helle; Langemets, Margit; Sepper, Maria-Maren; Argus, Reili. Tallinn: Eesti Keele Sihtasutus, 2009, 143-156.

Mihkla, M. (2007). Kõne ajalise struktuuri modelleerimine eestikeelsele tekst-kõne sünteesile – Modelling the temporal structure of speech for the Estonian text-to-speech synthesis., Tartu Ülikool) Tartu: Tartu Ülikooli Kirjastus.

Muischnek, K., Orav, H., Kaalep, H-J., Õim, H. (2003) Eesti keele tehnoloogilised ressursid ja vahendid: Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara. Tallinn: Eesti Keele Sihtasutus. ISBN 9985-79-053-7

Scott, M. (1997). WordSmith Tools version 2. Oxford: Oxford University Press. ISBN 0-19-459283-9.

Stubbs, M. (1996). Text and Corpus Analysis: Computer-Assisted Studies of Language and Culture. Oxford: Blackwell.

Veebipõhine kõnetuvastus. Loetud Internetis 12.märtsil 2011 aadressil http://www.phon.ioc.ee/dokuwiki/doku.php?id=projects:tuvastus:veebituvastus.et