Tallinna Ülikool

Informaatika Instituut

Kõrghariduskeele korpuse loomine

Seminaritöö

Autor: Kristo Kiis

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 kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on viidatud

....

(kuupäev) (autor)

Sisukord

Sisukord **3**

Sissejuhatus **4**

1. Ülevaade korpusest **6**
1.1. Mis on korpus? **6**
1.2. Suuremad korpused maailmas **6**
1.3 Korpuseid Eestis **9**

2. Korpuste võrdlus **11**
2.1 Korpuste võrldus **11**
2.2 Korpuste võrlduse kokkuvõte **17**
2.3 Näiteid erinevate korpuste otsingutulemustest ja otsingutest **18**

3. Korpuse disain **20**
3.1 Esitatud nõuded. **20**
3.1.1 Veebi lehe pool 20
3.1.2 Korpuse pool 21
3.2 Veebilehe nõute teostus **21**
3.3 Korpuse disain **23**
3.3.1 Korpuse kategooriateks jaotamine, nende administreerimine 24
3.3.2 Korpusesesse andmete sisestamine 24
3.3.3 Korpuses päring tegemine 24
3.3.4 Korpuses päringutulemuste kuvamine 25
3.4 Korpuse disaini kokkuvõte **25**

Kokkuvõte **27**

Kasutatud kirjandus **28**

Sissejuhatus

Käesolev töö on seotud “Erialakeeleõppe arendamisvajaduse uuringu” projektiga. Projekti eesmärgiks on aidata eesti keelt emakeelena mittekõnelevaid tudengeid õpetatavast paremini aru saama. Loengus kiirelt läbi töötatud tekstide ning napisõnaliste loengumaterjalide kasutamine võib osutuda tihti liiga keeruliseks ilma põhjalike eesti keele teadmisteta. Tänu sellele ei saa ilma tasemel erialase keele oskamiseta aineid täiemahuliselt omandada ja võib tekkida hariduse omandamist raskendavad lüngad. Samuti on vajalik erialase keele ja ainete täies mahus tundmine, et lõpetada ülikool ning asuda tööle Eestisse.

Eestis vene keelt kõnelevate abiturientide seas on eriti näha seda, et kõrghariduse omandamine jäetakse tahaplaanile või seda minnakse omandama välismaale kartuses, et erialane keel võib osutuda üle jõu käivaks. Hetkel on Tallinna Ülikooliski tehtud grupid neile, kelle emakeeleks pole eesti keel. Gruppideks on näiteks järgmised tasemed: B1, B2, C1 ja isegi A2. Ülikoolis õppimiseks peetakse piisavaks C1 taset. Kui aga tudeng alustab keeleõpet A2 tasemel, siis C1 tasemeni jõuab ta alles viimasel kursusel. Sinna vahele jääb peaaegu terve õppetöö, millest jääb tudengil palju omandamata tänu puudulikule arusaamisele materjalidest.

Projekti “Erialakeeleõppe arendamisvajaduse uuring” üheks etapiks on korpuse loomine ning sellega seoses on antud seminaritöö eesmärk toetada seda projekti vastava korpuse loomisega. Korpuse koduleht saab koosnema kahest etapist, üheks on kodulehe poole ettevalmistus ja ülesse seadmine, kooskõlas filoloogidega, kes on ka ühtlasi projekti autoriteks. Koduleht peab tulema võimalikult lihtne ning sisuhaldus väga kaasaegne ning paljude võimalustega.

Teise etapi osaks on korpuse loomine, mille tehniline teostus jääb bakalaureusetöö valmistamiseks. Seminaritöö käigus peab aga olema selge eesmärk ning tehniline sõnastus selle töö läbiviimiseks. Korpuse poolele hakkavad materjale lisama filoloogid ning samuti õppejõud. Õppejõudude kõnestiil, häältdus ja kõnetempo on küllaltki erinev, mis lisab korpuse keeleosale mitmekesisust. Korpusesse lisatud videoloengud annavad keeleõppijale võimalust harjutada keele tajumist ja sellest arusaamist. Tekstilised materjalid aitavad õppida konkreetse ainega seotud vajaliku sõnavara, mis on üheks eelduseks aine sooritamisel. Korpus hakkab mängima selles

projektis väga olulist rolli, kuna materjalide ning andmete vahetus hakkab käima just selle kaudu. Tulevikus on tudengitel koht, kust nad saavad arendada video- ning kirjamaterjalide teel oma erialast keeleoskust ja arusaamist sellest.

Esimeseks ülesandeks saab korpuse kodulehe panek serverisse. Selleks saab kasutada “wordpress” platvormi, mis on praegu väga populaarne ning hästi dokumenteeritud. Korpuse edasiarendamiseks tuleb sellele juurde programmeerida “lisavidinad”, mis lisavad sellele kodulehele korpusepoole. Kuna antud platvorm on hästidokumenteeritud, saab kasutada põhjana olemas olevaid platvormilisasid, mida ümberkohendades saab luua korpuse süsteemi. Samamoodi toetab “wordpress” erinevaid audio- ning videofailide üleslaadimisi ning esitamisi oma sisseehitatud flash-mängijaga, mis on korpuse loomisel oluline. Seminaritöö “lisavidina” arendamiseni kahjuks ei vii, kuid seletab lahti selle teoreetilise ning ülesehitusega seostuva poole, mis on olulised punktid arenduse alustamiseks.

Enne korpuse loomist tuleb uurida olemasolevaid korpuseid, et selgitada välja, mis on nende eelised ja ülesehitused. Teiste korpuste funktsionaalsust ja ülesehitust uurides saab ka uuele loodavale korpusele funktsionaalsust üle tuua. Samuti on seminaritöö eesmärgiks uurida korpuseid üldiselt ning koguda võimalikult palju infot tehnilise poole pealt.

1. Ülevaade korpusest

See peatükk käsitleb korpuse mõistet ning selle peamisi kategooriaid. Samuti on näidetena välja toodud osa korpuseid nii Eestist kui välismaalt

1.1. Mis on korpus?

Defineerides tänapäeva korpust*:* "*Korpuse all mõeldakse peamiselt polüfunktsionaalseid elektroonilisel kujul eksisteerivaid tekstikogusid, millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi kogu keelest* "(Muischnek, Orav, Kaalep, Õim, 2003).

Korpused jagunevad kaheks põhiliseks liigiks: kirjaliku keele korpused, mis koosnevad definitsiooni järgi tekstidest; ning suulise keele korpused. Suulise keele korpuste tähtsus on aegadega järjest tõusnud, eriti viimase 20 aasta jooksul(Stubbs, 2001). See on tänu sellele, et on avastatud suulise keele osatähtsuse kasvamist suuremaks kirjalikust, inimesed kirjutavad tekstides pigem seda, mida nad kuulevad kui grammatiliselt õigeid tekste.(Adolphs, 2008:19). Suulise keele korpused võivad sisaldada meedia faile, mis võivad olla nii audio kui video faili kujul. Kogu teksti osa suulise keele korpuses peab olema transkribeeritud kujul, mis annab rõhkudest ja kõnelejatest jms olulistest andmetest tekstis(Adolphs, 2008). Just suulise keele korpuse, mis sisaldab meedia faile loomise kirjeldus on selle seminaritöö üks eesmärke.

Korpuste ajaloost lühidalt rääkides olid juba 1950-ndatel aastatel need olemas ning ülesehituselt olid need lihtsad keelega seotud tekstikogud. Kaasaegne keelekorpus hakkas aga siiski arenema 80-ndate aastate algusest koos arvutitega, kuna arvutid kiirendasid korpustest tekstide otsingut ning nende kasutamine muutus seeläbi lihtsamaks ja populaarsemaks(Muischnek, Orav, Kaalep, Õim, 2003)

1.2. Suuremad korpused maailmas

Kõige suuremaks korpuseks maailmas on hetkel Korpora geschriebener Gegenwartssprache, milles asub 27.09.2011 seisuga 4,3 miljardit sõna. Korpuses olevate tekstide osas on nii ilukirjandust, teaduslikku kirjandust kui ka ajalehti ja teisi kirjalikke tekste. Korpus on kirjaliku saksa keele korpus, meediafaile seal ei ole. Korpusele puudub avalik ligipääs, andmeid saab tellida CD-plaadile või taotleda

ligipääsu serverile muul moel. Koostöös COMAS II IT agenguuriga on loodud otsinguaplikatsioon, mis võimaldab teostada otsingut selles andmebaasis. Otsingutarkvara on võimalik installeerida nii veebibrauserisse kui ka Windows'isse. (allikas: http://www.ids-mannheim.de/kl/projekte/korpora/)

Teine suuremate korpuste seas on ICE ehk *Internation Corpus of English.* See korpus loodi 1990-ndal aastal ning selle eesmärgiks oli koguda inglisekeelset materjali võrdlevate uuringute jaoks. Selles korpuses uurib 23 uuringurühma oma regioonioonile ja asukohale iseloomulikku inglise keelt. ICE alla kuulub ka USA suurim suulise keele korpus "Santa Barbara Corpus of spoken American". ICE korpuste uurimisrühmadel on kõigil üks eeskiri, mille järgi korpused on disainitud. Kodulehe andmetel on igas korpuses (suulises ja kirjalikus) umbes 500 teksti, kus on üle 2000 sõna ning kokku umbes miljon sõna. Tekste on kogutud alates 1990-ndast aastast. Autorid antud korpuses on täisealised ning on omandanud keskkoolihariduse inglise keeles. Samuti on autorid sündinud teatud regioonis või väga varakult kolinud sinna, et läbida kool inglise keeles. Tekstide autoriteks on erinevates vanusegruppides mehed ja naised. Selle suulise keele korpuse struktuur on välja toodud Tabel1 olevas tabelis.

Tabel 1

Suulise keele korpus (300)	Dialoogid (180)	Privaatsed (100)	Näost näkku vestlused (90) Telefonikõned (10)
		Avalikud (80)	Õppetunnid klassides (20) Eetris olevad vestlused (10) Eetris olevad intervjuud (10) Poliitilised debatid (10) Kohtu tunnistused (10) Äri tehingud (10)
	Monoloogid (120)	Skriptimata (70)	Spontaansed kommentaarid (20) Skriptimata kõned (30) Demonstratsioonid (10) Legaalsed presentatsioonid (10)
		Skriptitud (50)	Eetriuudised (20) (broadcast?)

			Eetrivestlused (20) Eetris mitteolevad vestlused (10)

Tabelis sulgudes olev number näitab tekstide arve selles kategoorias. Põhikategooriaks on suulise keele korpus, selle all on väiksemad kategooriad, milleks on dialoogid ning monoloogid. Dialoogid jagunevad kaheks: privaatsed ja avalikud. Privaatsed on sellised, mile õppematerjalid jm teaduslikud tekstid ei ole kõigile ligipääsetavad. Avalikud on aga need, mis on igale inimesele kättesaadavad keele harjutamiseks(Tabel1).(allikas: http://ice-corpora.net/ice/design.htm)

Välimuselt on ICE koduleht väga aegunud disaini ja lihtsa ülesehitusega, konkreetset baasi nende lehel ei ole, aga on näiteid riikide korpuste heli- ja tekstifailidest, mida saab lugeda ja kuulata. USA ja teiste maade korpuste faile saab alla laadida ning tellida CD-de ja DVD-na nende kodulehtedelt. Konkreetseid andmebaase, kust otsinguid teha, kahjuks ei leidnud. Kuna tegemist nii vana korpusega, siis sellepärast ka selline ülesehitus. Kaasaegsemaks saaks sellist suurt korpust teha pannes paika korraliku baasi, kuhu vanal kujul olevad andmed ümber kirjutada selleks, et saaks tulevikus lihtsalt lisada ning kuvada faile.

Järgmine korpus, mis mulle silma hakkas, on küll väiksem, aga sellegipoolest sisukam, kuna andmebaasis on olemas nii video- kui ka otsingupool, mille arendamise poole projekti raames loodav korpus püüdlebki. Projekt, mille raames korpus on loodud, kannab nime “SACODEYL”. See projekt tegeleb noorte kõne uurimisega järgmistes Euroopa suuremates riikides: Saksamaa, Prantsusmaa, Hispaania jt. Lindistatud on kooliõpilasi vanuses 13-18. Iga keele alamkorpus koosneb 20-25 lindistatud intervjuust õpilastega, pikkuses umbes 10 minutit.(allikas: http://www.um.es/sacodeyl/) Transkribeeritud tekstid asuvad selles korpuses xml-formaadis, kus on sees tekst, kõnelejad jm info video kohta. Failid, video- ja audiformaadis ning nende juures transkribeeritud tekstid on avalikult kättesaadavad. Selle korpuse otsing on väga põhjalik, sest saab otsida nii valdkonna kui ka grammatika järgi. Näiteks nii grammatika järgi minevikus aset leidnud tekstide kui ka erinevate eluteemade osas. Siinkohal tooks välja näiteks sellise teemajaotuse:

1. Teemad
    1.1 Isikukirjeldused

1.2 Kodu
1.3 Perekond
1.4 Igapäevategemised
1.5 Hobid (mis jaguneb veel teater, kino jms)
1.6 Loomad
1.7 Puhkus
1.8 Kool
1.9 Tulevikuplaanid

Selle korporatsiooni kategooriatesse jagamist saaks hästi rakendada ka meie loodavas korporatsioonis. Kuna lihtsalt teksti-, audio-, videokategooriatesse jagamisega võib tulla mõni nendest kategooriast väga pikk ning spetsiifilisema materjali otsimine ilma mingi otsingusõnata võib liialt aega võtta.

1.3 Korpuseid Eestis

Eestis suurimaks suulise keele korpuseks on Tartu Ülikooli suulise eesti keele korpus. Suulise keele korpus on mõeldud kasutamiseks ainult teaduslikul eesmärgil ning sellele tavakasutaja ligi ei pääse. See-eest on avalikuks kasutamiseks Eesti keele koondkorpus, mille eesmärgiks on kasvada 200 miljoni sõna suuruseks. Hetkel on see korpus juba väga põhjalik, sisaldades tekste aastast 1890. Samuti on seal ajalehe artikleid, teadusartikleid jms. Korpusele arendatud otsing on väga hästi välja töötatud, kuna see toetab regulaaravaldisi, mis on oluline spetsialistidele info pärimiseks, kuid sellel on ka natuke puudusi. Puuduseks on näiteks see, et täpitähed tuleb *html entitydena* kirjutada, mis näitab, et andmebaas pole utf-8 vaid on mingi kodeeringuga, mis ei toeta kõiki karaktereid. Samamoodi on otsing *case sensitive* , mis tähendab seda, et kui kirjutada “eesti” leiab eesti keelega seotud tekstid ja “Eesti” riigiga seotud tulemused. Kuigi sellistel otsingutel on tihti puudusi, kuna kasutajad võivad olla harjunud otsima andmeid ainult väiketähtedena. Kuigi kui see otsingule juurde kirjutada ei tohiks probleemi olla. (allikas: http://www.keeletehnoloogia.ee/projektid/koondkorpus)

Teiseks eesti keele korpuseks tooks välja Eesti Keele Instituudi korpuse, milles on 10,4 miljonit sõnavormi. Andmetest on aga 80% ajalehetekstid ning ülejäänud ilukirjanduslikud tekstid nagu näiteks: “Sõrmuste isand”, “Pipi Pikksukk” jt. Kõnekeele osa on selles korpuses väga väike, mõned osad seriaalist “Dallas”. Tekstid pärinevad selles korpuses ulatuvad kuni 2004. aastani. Otsingu osa on sellel korpusel koostatud

hästi, tulemusi kuvatakse lõikude kaupa ning otsingule leitud vasted värvitakse tulemustes ära. Hästi on ka lahendatud suur- ja väiketähe eristamise loogika otsingus, näiteks kui läbiva väike- või läbiva suurtähega kirjutades tuleb sama tulemus, kui aga algustäht suur, siis otsib sellisena nagu on. Samuti on Eesti Keele Instituudil ka suulise keele korpus, aga see on samuti teaduslikul eesmärgil ning piiratud ligipääsuga.

Eesti korpuste seas on ka väga väikseid korpuseid, kus otsingut ei ole vaja olnud korpusele, sest andmete hulk on nii väike, et saab kuvada listina ühe lehe peal. Sellisteks väikesteks korpusteks on näiteks: Uudistekorpus, kuhu on kogutud raadio uudised, mis asuvad ühes zip failis, mis sisaldab heli faile ning transkribeeritud tekste nendest failidest. Samuti on väga väike korpus loengu kõnekorpus, mis samuti sisaldab helifaile, aga lehe peal on need jaotatud valdkondadeks. On näiteks IT, matemaatika jt valdkondi. Loengu kõnekorpuse eesmärgiks on koguda 200 tundi loengumaterjale. Materjalide seas on nii Tehnika Ülikooli kui ka Tallinna Ülikooli loengumaterjale. Loengumaterjalid on wav-formaadis, mis on välja kujunenud põhiliseks korpuste audioformaadiks kuna kõik brauserid toetavad seda.

2. Korpuste võrdlus

Selles peatükis on välja toodud mõningad korpused, mille tehnilist poolt sai uuritud ja testitud. Võrldevateks kriteeriumiteks sai valitud korpuse tüüp, selle päring võimalused, programmeerimiskeel(kui võimalik tuvastada), korpuse suurus ning tulemuste kuvamine. Kriteeriumid on valitud lähtudes sellest, et tulemusi saaks kasutada käesoleva projekti jaoks korpuse loomiseks. Seminaritööle järgneb arendusuuring bakalaurusetöö tasemel, mille raames programmeeritakse korpusesse erinevaid tüüpi päringuvõimalusi. Antud seminaritöös on otstarbekas uurida kuidas päringuvõimalused on lahendatud teistes korpustes. Võrdluse jaoks olid valitud Eestis loodud korpused ning samuti rida välismaiseid korpusi, mis on suurema tõenäosusega populaarsemad. Oletus nende populaarsusest tuleneb sellest, et google-i otsinguga tulevad need kiiremini esile, mis viitab sellele, et nende poole sagedamini pöördutakse.

2.1 Korpuste võrldus

Järgnevalt on toodud korpuste analüüs, mis on paigutatud üksteise alla. Kriteeriumite nimetused on esile toodud rasvases kirjas.

1. **Korpuse nimi ja link:** Eesti vahekeele korpus (http://evkk.tlu.ee/Search)

Meediafailide või teksti korpus: Kirjakeelekorpus ehk tekstifailid

Päringuvõimalused: Otsing on väga spetsiifiline, st saab otsida autori soo, vanuse, emakeele jms järgi.

Programmeerimise keel: Zope raamistikul ja Python keeles.

Korpuse suurus: 513356 sõna. Korpusel on ka täpne statistika, kus on välja toodud kategooriate kaupa sõnade arv. Näiteks palju on meeste, palju naiste kategoorias sõnu.

Tulemuste kuvamine: Tulemusi kuvatakse lihtsas HTML-tabelis, kus päringu sisestatud sõna on kollase kirjaga välja toodud.

2. **Korpuse nimi ja link:** Eesti keele spontaanse kõne foneetiline korpus

www.murre.ut.ee/otsing/ekskfk.php

Meediafailide või teksti korpus: Teksti -ja meediafailide korpus.

Päringuvõimalused: Suhteliselt kitsad, saab kirjutada otsingusõna ning valida kolm kategooriat. Samas on huvitav see, et saab otsida mitte ainult sõna vaid ka rippmenüüst SAMPA või CV järgi. (Pilt1 täpsemaks infoks)

Programmeerimise keel: PHP ja MySQL.

Korpuse suurus: 20.12.2010 seisuga 179 554 sõna.

Tulemuste kuvamine: Siin kuvatakse tulemusi väga hästi, Tulemuste juures on helifailid olemas. Tulemuse esimeses reas on kirjas kontekst. Ning all helifail, kus see kontekst asub. Saab ka alla laadida otsingutulemuse tabeli(*textGrid*. Vt Pilt1

3. Korpuse nimi ja link: Eesti murde korpus www.murre.ut.ee/otsing/syntaks.php

Meediafailide või teksti korpus: Kirjakeele ehk tekstikorpus.

Päringuvõimalused: Otsinguparameetreid on vähe, saab valida kihelkonna, murde ning sõna kirjutada. Selle eest sõnasisestuse lahtris saab kasutada palju huvitavaid märgendusi, mis lihtsustavad otsimist. Näiteks saab kasutada .* ja *. Vastavalt millegi ees või järelt otsimisele, mis sarnaneb MySQL süntaksile. (Vaata Pilt2)
Programmeerimise keel: PHP ja MySQL.

Korpuse suurus: 03.08.2010 seisuga 1 153 165 sõna.

Tulemuste kuvamine: Tulemusi kuvatakse kontekstidena. Kontekstis on otsingusõna märgitud punasel taustal. Kontekstid on suhteliselt lühikesed.

4. Korpuse nimi ja link: Eesti emotsinaalse kõne korpus http://peeter.eki.ee:5000/reports/segments

Meediafailide või teksti korpus: Audiofailide korpus koos transkribeeritud tekstidega.

Päringuvõimalused: Otsing on huvitavalt koostatud, saab määrata tuvastusprotsendi.

Programmeerimise keel: Python ja PostreSQL.

Korpuse suurus: -

Tulemuste kuvamine: Tulemustena kuvatakse lauset ning linkidena heli- ja TextGrid faile, mis on transkribeeritud tekstid lausetest.

5. **Korpuse nimi ja link:** Keeleveebi otsingusüsteem http://www.keeleveeb.ee/

Meediafailide või teksti korpus: Teksti korpus (Teeb päringuid eesti keele koondkorpusest).

Päringuvõimalused: Otsing on väga lihtne, on ainult üks tekstiväli, millel puuduvad lisavõimalused.

Keeleveebi otsing tundub olevat lihtsustatud versioon sellest, mis on eesti keele koondkorpuse kodulehel.

Programmeerimise keel: PHP. Andmebaasi osas pole kindel, kuna päringud toimuvad teistest serveritest.

Korpuse suurus: Eesti keele koondkorpuse alusel 200 miljonit sõna.

Tulemuste kuvamine: On lihtsal kujul. On toodud välja laused, kus see sõna esineb. Ning värviga on välja toodud sõna esinemine lauses.

6. **Korpuse nimi ja link:** Vana kirjakeele korpus

http://www.murre.ut.ee/vakkur/Korpused/Otsi/mrg_paring.htm

Meediafailide või teksti korpus: Teksti korpus

Päringu võimalused: Otsing on jällegi huvitavalt lahendatud: saab sõnu Javascrip'iga juurde lisada, et otsitaks mitut sõna tekstis. Saab määrata veel sõnaliigi. Samuti on välja toodud erinevad kategooriad ehk kataloogid, mille järgi saab piirata otsingut. Näiteks kui vaja otsida sõna esinemist 16.saj tekstidest.

Programmeerimise keel: Ei ole kindel, tundub Python, kuna PHP ei ole kindlalt ning eelnevad selle projekti raames loodud korpused on ka Pythoniga.

Korpuse suurus: 700 000 sõna (korpus jaotatud kaheks kuni 18.saj ning 19.saj oma)

Tulemuste kuvamine: Tulemusi kuvatakse linkide listina, millel vajutades kuvab selle artikli / teksti, kus sõna(d) esines(id). Väga lihtsalt lahendatud.

7. **Korpuse nimi ja link:** British National Corpus http://corpus.byu.edu/bnc/

Meediafailide või teksti korpus: Teksti korpus.

Päringu võimalused: Otsing on lihtne, sõna sisestusväli ning sektsioonide valik(ajakiri vms).

Programmeerimise keel: .NET ja Microsoft SQL.

Korpuse suurus: umbes 100 miljonit sõna.

Tulemuste kuvamine: Otsitakse kahest baasist korraga, tulemused kuvatakse lihtsal kujul, tabelis linkidena. Peale vajutades avab uues aknas vastava teksti.

8. Korpuse nimi ja link: Russian National Corpus
http://www.ruscorpora.ru/en/search-main.html

Meediafailide või teksti korpus: Teksti korpus.

Päringuvõimalused: On kaks otsinguvarianti: lihtne otsing, kus ainult tekstiväli otsitava sõnaga ning laiendatud otsing. Laiendatud otsingus saab määrata sõna esinemise vormi, grammatilise vormi või käändevormi. Samuti saab otsida laiendatud otsingus mitme sõna esinemist korraga lauses või tekstis.

Programmeerimise keel: Pole teada.

Korpuse suurus: 150 miljonit sõna.

Tulemuste kuvamine: Tulemusi kuvatakse listina, kus on teksti pealkiri ning lause, kus see sees esineb. Kui lause on ainult mõnesõnaline on juures ka eelmine ning järgnev lause.

9. Korpuse nimi ja link: NoTa Oslo speech corpus http://www.tekstlab.uio.no/nota/oslo/english.html

Meediafailide või teksti korpus: On nii audio, video kui ka transkribeeritud tekstid.

Päringuvõimalused: Otsinguvõimalused on suhteliselt lihtsad, saab lisada mitut sõna korraga ning määrata sõna tüübi.

Programmeerimise keel: PHP ja MySQL

Korpuse suurus: umbes 900 000 sõna, tekstid on pärit 166 inimeselt, kes on sündinud ja kasvanud Oslos.

Tulemuste kuvamine: Tulemusi kuvatakse listina, kus alguses on lingid nii video- kui ka audiofailidele ning selle järel transkribeeritud teksti osa, kus sõna(d) esineb.

10. Korpuse nimi ja link: Corpus of historical american English http://corpus.byu.edu/coha/

Meediafailide või teksti korpus: Teksti korpus.

Päringuvõimalused: Otsing on lihtne, sõna sisestusväli ning saab valida aastakümnendi, kust seda otsida.

Programmeerimise keel: .NET

Korpuse suurus: umbes 400 miljonit sõna, alates aastast 1810

Tulemuste kuvamine: Tulemusi kuvatakse tabelis, grupeerib aastakümnete kaupa, kui valida üle ühe aastakümne. Samuti on link, mis viib täies mahus teksti juurde. Tabelis on ka välja toodud osa teksti, mis on enne leitud sõna ja peale leitud sõna.

11. Korpuse nimi ja link: Michigan Corpus of Academic Spoken English http://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;q1=word;page=simple

Meediafailide või teksti korpus: Teksti korpus

Päringuvõimalused: Otsingus saab sisestada sõna, ning määrata kategooria, Kategooriateks on näiteks autori sugu, valdkond, vanusegrupp jm. Väga süstematiseeritud. Saab ka leida kõige seast.

Programmeerimise keel: Tundub C keeles

Korpuse suurus: Ligikaudu 1,8 miljonit sõna

Tulemuste kuvamine: Saab valida, kas salvestada tulemused XML-formaadis või lihtsalt kuvada. Lihtsalt kuvades on sõna keskel ning selle ümber vasakul ja paremal on sellele eelnev tekst ja järgnev tekst, et oleks kontekstist aru saada.

12. Korpuse nimi ja link: Collection of Chinese corpora http://corpus.leeds.ac.uk/query-zh.html

Meediafailide või teksti korpus: Teksti korpus.

Päringuvõimalused: Saab sisestada sõna ning määrata, millisest korpusest otsida, toetab osaliselt regulaaravaldisi.

Programmeerimise keel: PERL

Korpuse suurus: üle 300 miljoni sõna.

Tulemuste kuvamine: Kuvatakse listina selliselt, et sõna on keskel ja sellele eelnev tekst (mis võib olla ka poolik lause) ning järgnev tekst on samal real. Eelnev ja järgnev tekst on sõnade arvu järgi lõigatud, et otsingutulemused mahuks ühele reale.

13. Korpuse nimi ja link: IPI PAN korpus (Poola) http://korpus.pl/poliqarp/poliqarp.php

Meediafailide või teksti korpus: Teksti failide korpus

Päringuvõimalused: Sõna sisestusväli, millele järgnevad sorteerimise ning tulemuste kuvamise seaded.

Programmeerimise keel: PHP – siiani kõige kaasaegsem ja ilusam korpuselahendus, kust palju võtta. Disain ja kasutaja liides on ilusasti tehtud lähtudes Web 2 standartitest.

Korpuse suurus: ligikaudu 250 miljonit sõna.

Tulemuste kuvamine: Tulemuste kuvamine on hästi lahendatud, tuleb list lausetest, kus see sõna sees on. Lausele peale vajutades kuvab lehe all osasse tervik teksti.

14. Korpuse nimi ja link: Corpus of the Contemporary Lithuanian Language http://donelaitis.vdu.lt/main_en.php?id=4&nr=1_1_2&kalba=en

Meediafailide või teksti korpus: Teksti korpus.

Päringuvõimalused: Lihtne otsingusüsteem, kus peale sõna kirjutamise saab ka valida, mis allikast see on, kas ajalehest, raamatust vms.

Programmeerimise keel: PHP

Korpuse suurus: ligikaudu 102 miljonit sõna

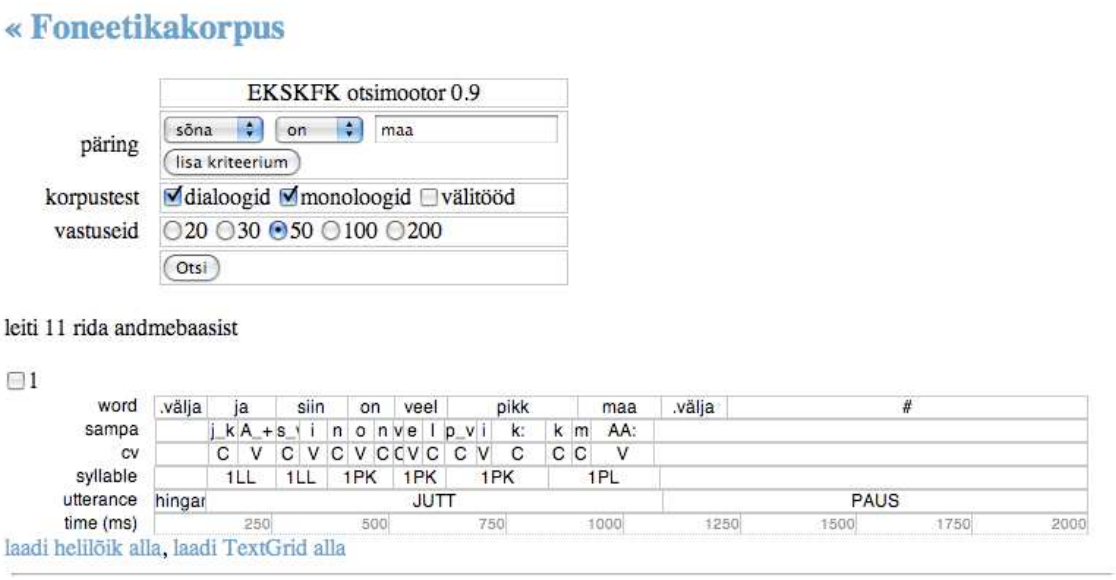
Tulemuste kuvamine: Tulemusi kuvatakse tabelis kategooriate kaupa, kui on valitud mitu. Kategooriale vajutades kuvab listina kontekstid, kus see sõna esineb.

2.2 Korpuste võrlduse kokkuvõte

Kokkuvõtteks võrdluse kohta oli korpuseid erinevates suurustes, suurus sõltus korpuse eesmärgist ning aktiivsusest ja ka materjalide kätte saadavusest. Eesti suurimaid korpuseid saab võrrelda välismaa korpuste suurusega, sõnade arvus, mis ulatub miljonitesse. Väiksemad aga sadadesse tuhandetesse. Meediafailide arvu poolest olid korpused ühtlasemad. Paarsada tundi materjale tundub olevat keskmine sellise korpuse suurus. Eesti uurimise all olevaid korpuseid oli 6 ning need kõik said võetud http://www.keeletehnoloogia.ee/ lehelt, kus on mitmeid käimas olevaid ja lõppenud keelealaseid projekte. Välismaa korpuseid sai leitud Google-i otsingumootori järgi ja kokku väljatoodud 8 huvitavamat, millel andis selgelt uurida antud kriteeriume.

Tehnilise poole pealt sai palju huvitavaid lähenemisi otsingu jaoks ning tulemuste kuvamise jaoks. Kõigil välja toodud korpustel olid otsingud olemas ja neis oli läbiv süsteem, et on otsingusõna ning kataloogid, kust otsida. Tulemuste osas oli samuti läbiv teema tulemuse esinemine lauses värvida või *bold* –iks teha. Samuti lausele, kus sees sõna esines toodi välja ka sellele eelnevat teksti kui ka järgnevat teksti, arvestades realaiusega. Korpusi oli erinevates programmeerimiskeeltes, tuli välja ka muster. Vanemad korpused on PERL ja C keeles, uuemad aga Pyhton ja PHP keeles. Tavaliselt sama projekti raames loodud korpused said samade arendusmeetoditega tehtud.

2.3 Näiteid erinevate korpuste otsingutulemustest ja otsingutest



Pilt 1. Foneetika korpuse otsingu vorm(www.murre.ut.ee/otsing/ekskfk.php)

Päringu osas saab valida erinevate korpuste vahel, kus otsing toimub. Tulemuste osas on tabeline välja toodud lause ning transkribeeritud kujul tekst, kus on näidatud heli faili pikkus ning mis sõna lausest tol ajahetkel esineb. Samuti alla laadimis võimalused heli ja transkribeeritud teksti failil. (Pilt1)

kontekst	fail (:sõne nr)
linad üleväl=ja (...) ja sa- sara all õl'i käsi mas'sin kellega	Idamurre/KOD_Miili_Lepp_F0013_synt.txt (:313)

näidetakse 1. lehekülge 1-st. (leiti 1 vastust)

Pilt 2. Murdekeele korpus (www.murre.ut.ee/otsing/syntaks.php)

Päringu osas on sõna sisestus ja kihelkonna valik, saab ka mitu valida ctrl klahviga. Kontekst on mitut sõna kuvab tulemustes. Tulemustes kuvab konteksti ja ka faili asukoha allalaadimiseks(Pilt 2).

3. Korpuse disain

3.1 Esitatud nõuded.

Järgnevates punktides on välja toodud projekti autorite esitatud nõuded nii kodulehele, mis valmib seminaritöö käigus kui ka korpusele, mille arendus toimub kodulehe “vidinana” bakalaurusetöö etapis. Nõuded olid esitatud projektis kaasatavate filologide poolt.

3.1.1 Veebi lehe pool

Eesmärgid:

- lühidalt tutvustada EKAV projekti eesti ja vene keeles (venekeelne tutvustus võiks olla lingitud ka TLÜ Katariina Kolledžiga)
- mahutada ja säilitada nii suulise kui ka kirjaliku korpuse infot (andmete üleslaadimine peaks olema kiire ja käepärane. Kogu info võiks paikneda TLÜ serveris)
- nii üliõpilaste kui ka õppejõudude ankeedid peaksid olema täidetavad e-Formularis ja täidetud ankeedid võiksid säilida ka TLÜ serveris piiramatu aja jooksul) Samas võiksid täidetud ankeedid olla kättesaadavad ja loetavad vaid projekti täitjatel.
- eraldi lingina tuleks välja tuua konfidentsiaalsust ning andmekaitset tutvustavat infot
- projekti käigus peaks olema võimalus pidevalt kodulehte täiendada (kõikidel projekti täitjatel peaks olema võimalus seda teha), nt võiks olla eraldi link valminud publikatsioonidele ja lõpuüritusele

Sisupunktid ja struktuur:

- esilehel sissejuhatav jutt, mis tutvustab projekti olemust nii eesti kui ka vene keeles
- e-Formularis ankeedid (õppejõududele eraldi ja tudengitele eraldi), täitmisel võiks kindlasti olla viide (nn süvalink) konfidentsiaalsus- ja andmekaitse memodele

- link kirjalikule korpusele ja suulisele korpusele
- tulevaste publikatsioonide koht
- info teavitusürituse kohta

3.1.2 Korpuse pool

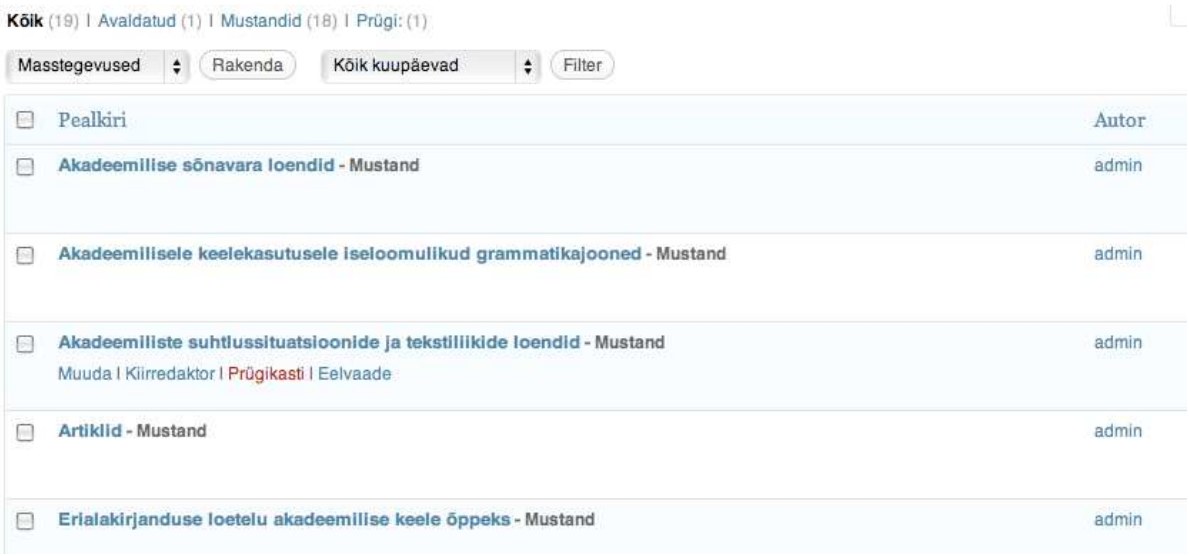
Eesmärgid:

- Tulevikus kasutatakse korpust võimalikult laialt erialakeeleloengute ettevalmistamisel ja läbiviimisel ja erialakeele uurimisel – nii toormaterjali (videod, audio), litereeringuid kui annoteeritud litereeringuid.
- Videod, audiod, märgendamata litereeringud ja märgendatud litereeringud (nende segmendid) peavad kõik olema üksteisest lahus kasutatavad (tuleb arvestada, et tarkvara, annotatsioonistandardid võivad aeguda, aga videod ja nende mahakirjutused võiksid jääda!)
- Kättesaadavad peaks olema täisvideod, ortograafiline märgendus ja kollektsioon videojuppe, mille juures on täpne info, kust nad pärit on.
- Salvestus peaks tulema kokku 11,5 h ehk 690 min, nt keskmiselt 3minutilisi juppe võiks olla u 230 tk (8 erinevat loengut/situatsiooni, igaühest u 30 juppi).
- Õppeotstarbel subtiitrid? Selle võiks IT-spetisalist meie ortograafilise litereeringu pealt otse võtta.
- Väga oluline on videote-videolõikude kasutusmugavus, eriti keeleõpetaja jaoks. Et saaks otse Mozillas või Windows media playeriga või VLC playeriga (see on vabavaraline) vaadata, et poleks vaja programme alla laadida.
- kirjaliku korpuse juures võiks olla koht, kuhu projektist osavõtvad õppejõud saaksid oma materjalid lisada (võiks olla võimalik nii doc.,docs., kui ka text.-formaadis kirjutisi lisada)

3.2 Veebilehe nõute teostus

Peale veebilehe ning korpuse nõuete lugemist tuli teha otsus, kas luua mingi enda tehtud süsteem või leida olemasolev. Kuna tegemist ei ole ainult korpuse poolega vaid peab ka olema toimiv ja lihtsasti hallatav koduleht, tuli otsus, et peaks olema olemasolev ning pidevalt uuenev ja kaasaegne sisuhaldussüsteem. Hetkel tundub selleks olevat Wordpress, mis on pidevalt uuenev ning “vidinate” poolest pidevalt täienev.

Edasiseks etapiks sai välja uuritud Wordpressi nõuded serverile, mis sai edastatud Tallinna Ülikooli Informaatika Instituudi süsteemi administraatorile, kelle käest sai tellitud serveriruum koos MySQL andmebaasi loomise õigustega. Peale serveri andmete kättesaamist sai installeeritud Wordpress antud keskkonda aadressile: http://minitorn.cs.tlu.ee/~kiis , kus see ka püsib kuni arenduse lõpuni. Hetkel on üleval süsteem vajalike "vidinatega", kodulehe osa jaoks on olemas keelte süsteemi vidin, mis võimaldab mitmekeelsust kodulehel, kui ka e-formularide jaoks lisatud vidin: "*contact form 7*", mis on kõige populaarsem Wordpressi vormi vidin. Ülejäänud kodulehe funktsionaalsuse saab täita juba olemas olevate Wordpressi võimalustega.



Pilt 3. Osa filoloogide sisestatud menüüpunkte ja tekstilehti Wordpressis (http://minitorn.cs.tlu.ee/~kiis/wp-admin/edit.php?post_type=page)

Pildil on näha juba filoloogide poolt lisatud menüü punkte, mida on hetkel paari kümne ringis. Musta kirjaga on näidatud hetke staatus menüü punktil ning autor admin(Pilt3)

Lisaks kodulehel sai vahetatud ka teema disain, mis on sobilikum korpuse jaoks, kui alguses automaatselt kaasa tulev disain, aga korpuse etapis võib see muutuda, sõltuvalt korpuse vajadustest.

Pilt 4. Kodulehe disain ja avalik vaade.

Pildil on näha on hetke disaini lehel, kus on avalik ainult 1 menüü punkt ning lihtne otsingu väli. Disain on lihtne, värve ei ole kastuatud, et hoida neutraalsust(Pilt 4).

3.3 Korpuse disain

Tuleks luua uus vidin, kuna olemasolev Wordpressi sisulehtede moodul ei toeta nii paindliku süsteemi korpuste materjalide hoidmiseks. Wordpressi sisulehtede näol küll saaks üles ehitada korpuse struktuuri - need toetavad ka meediafaile ning õigusi, aga ei oleks otstarbekas hoida kodulehe sisu koos korpusesisuga ühtse moodulina. Selline erinevate ülesannetega andmete kooshoidmisega võib tulla korpuse kasvamisel palju segadust, näiteks mõni tekst võib sattuda sisusse ja sisutekst korpusesse. Samuti oleks problemaatiline hea otsingu tegemine, sest otsingu tulemustes peaks ilmuma ainult septsiifilised korpusematerjalid, mitte sisutekstid. Nendest probleemidest hoidumise vastu aitabki eraldiseisev vidin, mis kasutaks andmebaasis oma tabelit, mida saaks ise valmistada vajaduse järgi. Näiteks teksti sisestamisel saaks määrata kategooriat, teksti autorit, sisestajat jms. Samuti “vidina” kasutamine lihtsustab administraatoritel ja sisestajatel paremini navigeerida materjalides. Järgnevalt jagame korpuse loomise eraldi ülesanneteks, mida tuleks arvestada selle arendamisel ning ülesannetele võimalikud lahendused, kas uuritud korpuste alusel või projekti autorite soovitusel.

3.3.1 Korpuse kategooriateks jaotamine, nende administreerimine

Kategooriaid saaks jagada tüübi kaupa või valdkonna kaupa. Esialgu on projekti autoritega kokku lepitud sisestava kirje tüübi järgi, kas tegu on audio-, video- või tekstifailiga. Korpuse kasvamisel peaks arvestama ka muude kategooriate võimalustega.

3.3.2 Korpusesesse andmete sisestamine

Andmeid peaks saama korpusesse sisestada lihtsalt. Näiteks valida kategooria ja edasi sisestava kirje tüüp, kas audio-, video- või teksti fail. Kui kategooriad on tüübi järgi siis koonduvad need kaks valikut üheks. Edasi tuleks sisestada teksti- või meediafaili pealkiri. Mõtekas oleks meedia- ja tekstifailide jaoks eraldi vormid teha, kuna väljade nimetused võivad muutuda ning olenevalt sisestavast tekstist või meediast võib mingi väli segadusse ajada. Näiteks teksti kirjel transkribeeritud teksti lisamine. Kindlasti peaks olema autori sisestamine ning idee oleks ka Javascriptiga jälgida sõnade arvu teksti sisestamisel või meedia sisestamisel transkribeeritud teksti sõnade arvu. Sõnade arvu võiks jälgida lihtsa Javascripti funktsiooniga, mis arvestaks tühikuid tekstis. Kuna uuritud korpustel on peaaegu kõigil välja toodud sõnade arv, lihtsustaks see sõnade arvu arvutamist lehe administraatoril. Meediafailide sisestamisel saaks kasutada “vidina” kirjutamisel olemasolevaid Wordpressi meediafailide sisestamisfunktsioone, mis tundusid uurimisel väga hästi tehtud olevat.

Andmete sisestamisel tuleb arvestada ka, et nende sisestajad, kes ei pruugi olla lehe administraatorid vaid hoopis õppejõud peaksid saama sisestada ka dokumendi faile, näiteks doc või pdf. See eeldab, et vidin peaks saama kätte teksti põhilistest dokumendi failidest, et otsing saaks otsida nende tekstide seas. Siin saaks läheneda kahte moodi, kas otsida välja vastavad class-id mis oskavad teksti välja lugeda dokumentidest või lasta sisestajatel ka copyda tekst eraldiseisvalt teksti välja.

3.3.3 Korpuses päring tegemine

Korpuses päringute tegemise jaoks peab olema “lisavidinana” arendatud hea otsingumootor, mis toetaks erinevatest kategooriatest otsimist ning kas regulaaravaldisi või muid mitme sõna või sõna sõna seest otsimismeetodeid. Elementaarseteks ning kõige levinumateks otsingus kasutatavaks süntaksiks uurimuse käigus ostutus sõna.* või *. sõna kasutamine, mis tähendab, et otsing otsib seda sõna ka sõnade seest vastavalt kas algusest või lõpust. Samuti võiks otsingus olla ka sorteerimiskriteeriumid.

Elementaarseks oleks uuemad sisestused eespool. Otsing peab ka arvestama tulemuste õiguseid, see tähendab, et kas antud kasutajal on õigus kõikidele materjalidele või ainult temale piiratud hulgal materjalidele. Näiteks sisselogimata kasutaja saab ainult avalikke andmeid vaadata, aga sisseloginud kasutaja näeb lisaks ka mitteavalikke materjale otsingutulemuste seas.

3.3.4 Korpuses päringutulemuste kuvamine

Päringutulemusi hakkab otsing kuvama listina, kus oleks välja toodud laused, kus on sees otsitav sõna. Laused on pärit tekstifailide puhul tekstidest ning meediafailide puhul transkribeeritud tekstist. Lause sees seisab märgitud sõna rasvase kirjaga, et eristada seda sõna lauses. Meediafaililiste tulemuste puhul on lisaks tekstile ka lingid meediafailile. Esialgu on kokku lepitud, et meediafailid on .flv ja .wav kujul, mis võimaldab kasutada olemasolevat Wordpressi meediafailide esitajat. Audiofaile suudab brauser ise esitada, aga videofailide jaoks peab kasutama kindlasti Wordpressi meediamängijat.

Tekstiliste tulemuste puhul on link, mis avab allpool otsingutulemusi selle teksti ja märgib lehe fookuse sellele. Fookus on hea, kui leht läheb liiga pikaks, et kasutajal oleks parem navigeerida. Kindlasti peab lehe all olema ka link lehe ülesse, et kui kasutaja ei ole tulemustega rahul, saab ta ilma kerimata lehel üleval olevaid otsinguparameetreid muuta. Samuti peab ka arvestama, et kui tekst on välja võetud dokumendifailist võib seal esineda puudusi ja peaks olema juures ka originaaldokumendi fail, allalaadimiseks.

3.4 Korpuse disaini kokkuvõte

Kokkuvõtteks on kodulehe pool väga lihtne, mis on juba valmis ning projekti autoritele kättesaadav. Hetkel saavad nemad ka toimetada kodulehe struktuuri ja tekste. Kodulehe ülesehituses võib tulla bakalaurusetöö osas muudatusi, kuna korpuse osa tekkimisel tuleb umber mõelda struktuuri, aga põhisüsteem jääb samaks. Wordpress on väga paindlik platvorm ning omab kõiki funktsionaalsusi tavalise kodulehe toimetamiseks ja seadmiseks. Samuti on olemas ka “vidinad”, mida läheb võib-olla vaja kodulehe edasiarendamisel. Dokumentatsioonist on väga lihtne välja lugeda, mis moodi ”vidinaid” luua ja olemasolevaid funktsionaalsusi ümber kirjutada. Näiteks kasutajagruppide lisamine ja otsingu täiendamine ja arendamine. Praeguse seisuga on

Wordpress ostunud heaks valikuks. Samuti täidab Wordpress kõik kodulehe osa jaoks püstitatud eesmärgid ning erinevate brauserite nõuded.

Kokkuvõte

Seminaritöö eesmärgiks oli luua lihtne veebileht projekti autoritele ehk filoloogidele korpuse kodulehe jaoks. Veebileht sai ehitatud wordpressi peale, mis on pidevalt uuenev sisuhalduse mootor, kül blogimise jaoks, aga väga edukalt kasutatav ka väiksemate kodulehtede jaoks. Veebileht sai loodud edukalt ning filoloogid saavad hästi hakkama sisuhaldamise ning uue sisu ehitamisega sinna peale. Selle osa saab lugeda edukaks.

Teiseks suuremaks eesmärgiks oli seminaritöös uurida korpuse loomise võimalusid ning selle sidumist kodulehega. Uurides Wordpressi võimalusi selgus, et vidinana on seda hea luua Wordpressi peale. Teiste korpuste uurimisest selgus väga hästi, et milline peaks korpus välja nägema ning kuidas peaks toimuma andmete sisestamine ja nende kuvamine veebilehel. Andmete sisestamist ja kuvamis osa sai läbi arutatud filoloogidega ning teostus idee sai paika panna lähtudes nende esitatud nõuetest. Samamoodi sai veebileht teostatud nõuete järgi.

Seminaritöö eesmärk oli ka ettevalmistus bakalauruse töö teostamiseks. Sai ka täidetud see ülesanne edukalt. Kuna bakalauruse töö eeldab teadmisi korpustest ja ka teadmisi nende ülesehitusest ja funktsioonidest sai seminaritöö hästi ülesse ehitatud, mis annab selge ülevaate korpustest ja ülevaate mingist hulgast.

Kokkuvõttes said eesmärgid täidetud ning Seminaritöö faasis on korpuse koduleht töö valmis ning ootamas korpuse osa arendust.

Kasutatud kirjandus

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

Adolphs, S. 2008, *Corpus and Context, Investigating pragmatic functions in spoken discourse*. John Benjamin publishing company.

Eesti keele instituudi korpus. Loetud internetis 11. oktoobril 2011, aadressil http://www.eki.ee/corpus/

Eesti keele koondkorpus. Loetud internetis 12.oktoobril, aadressil http://www.keeletehnoloogia.ee/projektid/koondkorpus

The design of ICE corpora. Loetud internetis 11. oktoobril 2011, aadressil http://ice-corpora.net/ice/design.htm

Käimasolevad EKKTT projektid. Loetud internetis 12. oktoobril 2011 aadressil http://www.keeletehnoloogia.ee/projektid

Korpora geschriebener Gegenwartssprache. Loetud Internetis 10. oktoobril 2011 aadressil http://www.ids-mannheim.de/kl/projekte/korpora/

Muischnek, K., Orav, H., Kaalep, H-J., Õim, H. (2003) Eesti keele tehnoloogilised ressursid ja vahendid: Arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara.

Stubbs, M. 2001, *Words and Phrases: Corpus studies of Lexical Semantics*. Oxford: Blackwell

SACODEYL corpora. Loetud internetis 11. oktoobril 2011, aadressil http://www.um.es/sacodeyl/

Korpuse üldiseloomustus. Loetud internetis 13. oktoobril 2011, aadressil http://www.cl.ut.ee/suuline/Ylevaade