Tallinna Ülikool
Digitehnoloogiate Instituut

Tekstide võrdlemine eesti vahekeele korpuses

Seminaritöö

Autor: Hanno Rudissaar

Juhendaja: Jaagup Kippar

Autor: ..... „ ..... „2016

Juhendaja: ..... „ ..... „2016

Instituudi direktor: ..... „ ..... „2016

Tallinn 2016

Autorideklaratsioon

Deklareerin, et käesolev seminaritöö 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

Sissejuhatus

Teema on aktuaalne keeleteadlastele, kellele sõnasageduste teadasaamine tekstis on vajalik.

Idee teha Eesti vahekeele korpusele(edaspidi korpus) täiendus tuli kiiruga. Valitud sai suvaline saadaolev variant, mis tundus huvitav ja kasulik, see tähendab, et tulemusest on kellelegi reaalselt kasu.

Töö eesmärgiks on luua tekstide võrdluse funktsionaalsus eesti vahekeele korpusele. Funktioon, kui selline, on abiks keeleteadlastele tekstied kohta informatsiooni saamiseks. Samuti on sellest kasu ka korpuse kasutajale endale, saades teada enda tekstis olevate sõnade sageduse ja muud informatsiooni.

Töös on juttu kuidas Zope sai seadistatud ning mis probleemid sellega kaasenesid. Samuti koodifailide asetsemine serveris ning ka erinevad näited tekstivõrdluse kohta.

1. Eesti vahekeele korpus

Juttu tuleb sellest, mis on eesti vahekeele korpus ning mida sinna juurde vaja on.

1.1 Mis on korpus?

Tallinna Ülikooli eesti vahekeele korpus (EVKK) on eesti keele kui riigikeele (teise keele) ja võõrkeele õppijate kirjalike tekstide kogu. EVKK-s on rida alamkorpusi, kasutajaliides, mitmetasandiline annoteerimis- ja märgendussüsteem, statistikamoodul, tekstide automaatanalüüsi võimalused jm. Kombineerides erinevaid alamkorpusi, tekstilisi tunnuseid, vealiike ja metateavet õppija kohta, võimaldab korpuse kasutajaliides teostada mitmetasandilist otsingut.

Korpust saab kasutada empiirilises ja rakenduslikku laadi uurimistöös; tulevaste õpetajate ja lingvistide koolitamisel; tegevõpetajate täiendõppes jm. [1]

Rohkem saab lugeda korpuse kohta Virgo Halliku seminaritööst[5] või Siim Medijainen kaitstud diplomitöö esimesest peatükist [6].

1.2 Vajaminevad tööd

Luua tuleb funktsioonid, mis võimaldavad tekstide kohta informatsiooni saada, nagu sõnade ja lausete arv, mis on keskmine sõnapikkus, erinevate sõnadearvuga lausete osakaal kogu tekstist, mitmesõnaliste sõnade osakaal ning lõpuks ka võimalus võrrelda või näha oma teksti erinevust varem hinnatud tekstidega.

2. Tutvumine korpusega

2.1 Probleemid

Enne kui sai korpuse kallal töötama asuda, tuli kõigepealt mõningased probleemid lahendada. Kuna tegu oli mahukate ja tähtsate failidega, siis kogu tegevus käis üksikute failidega. Sellega seoses tekkis esimene probleem – hakati erinevate muudatustega ühte ja sama faili üle kirjutama. Kogu töö mis vahepeal tehtud sai, oli tühine, see kustus. Lahendus sellele oli tegelikult suht lihtne, leppisime kokku, kes millise failiga hetkel tegeleb ja probleem lahenes. See ei olnud ainuke murekoht. Veel tekkis probleem Zope eripäraga, aegajalt tekkis nö „koodiviga“. Pikalt sai uuritud, et mis koodis ikkagi valesti on ja selgus, et tabulaator ei ole aktsepteeritav. Zope nõudis nelja tühikut. Kuna Notepad++ polnud vastavalt seadistatud tekkis sellega seoses probleeme edaspidigi.

2.2 Koodifailid

Kogu korpus paikneb greeny serveris, greeny.cs.tlu.ee. Greeny-s on mitu erinevat testimiseks mõeldud korpust. Konkreetne töö valmis testzope6-s „/bkp/testzope6/Products/Korpus“ . Kogu süsteem on ülesehitatud loogiliselt, tähtsamad koodijupid, millega autorgi tegeles, on „Korpus“ kaustas ning veebilehitseja mallid - veebilehe disain, asuvad kaustas „/bkp/testzope6/Products/Korpus/browser/templates/www“, kus sai muudetud „usertext.pt“ faili.

3. Arendustöö

3.1 Algus

Töö algas lihtsamate funktsioonide töölesaamisega. Algselt sai kokku loetud, mitu sõna või lauset kasutaja tekstis olemas on. Seejärel detailsemalt, keskmine sõnapikkus, samuti ka lühim ja pikima sõna sõnapikkus. Veel, mis sõna on kõige lühem või pikem ning sama tulemus ka listina, kui tulemusi oli rohkem kui üks. Seejärel tuli välja mõelda moodus, kuidas võrrelda kasutaja enda teksti olemasolevate ja hinnatud tekstidega. Sellest tuleb juttu eraldi paragrahvis (Joonis 3).

Teksti üldandmed							
Sõnade arv	**Lausete arv**	**Keskmine sõnapikkus**	**Lause keskmine pikkus**	**Pikima sõna pikkus**	**Pikimad sõnad**	**Lühima sõna pikkus**	**Lühimad sõnad**
240	23	7.09	10.43	25	['sotsialiseerumisprotsessi']	0	['']

Joonis 1. Esialgsed näited.

3.2 Uuritava teksti keeletaseme leidmine

Suvepraktika käigus tuli mõelda välja variant, kuidas hinnata kasutaja teksti varem keeleteadlaste poolt hinnatud tekstidega. Kuna pärida saab keeletasemete kaupa, siis tuli välja mõelda omadused mille alusel teksti võrdlema saaks hakata. Lõime koefitsendid, mille alusel tekst saab tunnused, mille alusel võrreldakse. Erinevus leitakse eelnevalt hinnatud teksti ja kasutaja enda teksti tunnuste alusel. Erinevuse arvutamine on lihtne, lahutatakse hinnatud teksti omadus kasutaja teksti omadusega ja võttes arvesse valituid parameetreid. Kasutaja ise määrab, millist tunnust tuleks arvestada ja kui palju. Kui koefitsent on 0.0, siis seda ei arvestata üldse, kui 1.0 siis täielikult. Koefitsendi parameetriteks valisin sõnade arv tekstis, lausete arv, keskmine sõnapikkus, keskmine lühim sõna, keskmine pikim sõna, kahe-, kolme-, nelja-, viie-, 6...9- ja 10...20 – täheliste sõnade protsent ning kahe-, kolme-,nelja-, viie-, 6...9- ja 10...20-sõnaliste lausete protsent vastavas tekstis. Kõiki parameetreid saab kasutaja käsitsi sättida, vaikeväärtused parameetritel on seatud järgmiselt: Sõnadearv tekstis on 0.5, lausetearv 0.5, keskmine sõnapikkus 0.0, keskmine lausepikkus 1.0, keskmine lühim sõna 0.0, keskmine pikim sõna 0.5, kahetäheliste sõnade protsent 0.5, kolme-, nelja-, viietäheliste sõnade protsent 0.0, 6...9-täheliste sõnade protsent 0.5, 10...20-sõnaliste lausete protsent 1.0, kahesõnaliste lausete protesnt 0.5, kolme-, nelja-, viiesõnaliste lausete protsent 0.0, 6...9-sõnaliste lauste protsent 0.5, 10...20-sõnaliste lausete protsent 1.0. (Joonis 1)

```
'sonadearv': 0.5,
'lausetearv': 0.5,
'keskminesonapikkus': 0.0,
'keskminelausepikkus':1.0,
'keskminelyhimsona': 0.0,
'keskminepikimsona': 0.5,
'2wordpercent': 0.5,
'3wordpercent': 0.0,
'4wordpercent': 0.0,
'5wordpercent': 0.0,
'6to9wordpercent': 0.5,
'10to20wordpercent': 1.0,
'2sentencespercent': 0.5,
'3sentencespercent': 0.0,
'4sentencespercent': 0.0,
'5sentencespercent': 0.0,
'6to9sentencespercent': 0.5,
'10to20sentencespercent': 1.0
```

Joonis 2. Parameetrite väärtused.

Olenevalt parameetrist tuli mõelda välja viis, kuidas see endale väärtuse saab. Sõnade arv ja lausete pikkuste koefitsendid jagatakse 100-ga, sõnade ja lausete arvu parameeter arvutatakse kasutades logaritmi alusel 10. Seda seepärast, et kui mingis tekstis peaks sõnu või lauseid olema oluliselt rohkem, siis erinevus ei muutuks tohutult(Joonis 2).

```
435     erinevus=\
436       log(sonadearv, 10)*koef['sonadearv']+\
437       log(lausetearv, 10)*koef['lausetearv']+\
438       abs(vordlus['averagewordlength']-keskminesonapikkus)*koef['keskminesonapikkus']+\
439       abs((vordlus['wordcount']/vordlus['sentencecount'])-keskminelausepikkus)*koef['keskminelausepikkus']+\
440       abs(vordlus['averageshortestword']-keskminelyhimsona)*koef['keskminelyhimsona']+\
441       abs(vordlus['averagelongestword']-keskminepikimsona)*koef['keskminepikimsona']+\
442       abs(vordlus['2wordpercent']-word2percent)*koef['2wordpercent']/100.0+\
443       abs(vordlus['3wordpercent']-word3percent)*koef['3wordpercent']/100.0+\
444       abs(vordlus['4wordpercent']-word4percent)*koef['4wordpercent']/100.0+\
445       abs(vordlus['5wordpercent']-word5percent)*koef['5wordpercent']/100.0+\
446       abs(vordlus['6to9wordpercent']-word6to9percent)*koef['6to9wordpercent']/100.0+\
447       abs(vordlus['10to20wordpercent']-word10to20percent)*koef['10to20wordpercent']/100.0+\
448       abs(vordlus['2sentencespercent']-sentences2percent)*koef['2sentencespercent']/100.0+\
449       abs(vordlus['3sentencespercent']-sentences3percent)*koef['3sentencespercent']/100.0+\
450       abs(vordlus['4sentencespercent']-sentences4percent)*koef['4sentencespercent']/100.0+\
451       abs(vordlus['5sentencespercent']-sentences5percent)*koef['5sentencespercent']/100.0+\
452       abs(vordlus['6to9sentencespercent']-sentences6to9percent)*koef['6to9sentencespercent']/100.0+\
453       abs(vordlus['10to20sentencespercent']-sentences10to19percent)*koef['10to20sentencespercent']/100.0
```

Joonis 3. Erinevuse arvutamine koodina.

3.3 Päringud teksti omaduste kohta

Tekstitasemete erinevuse arvutamiseks tuli luua baas, mille alusel võrdlust tehakse. Otsustati, et korpuse päringud tuleks salvestada, et need võrdlemise aluseks võtta. Päringute salvestamisel salvestatakse kõik informatsioon, mida pärida saab, nagu mis sõna otsiti, mis korpusest tekst pärit on, kus tekstikirjutaja elab, tema vanus, tema emakeel, tema keeletase, kas ta kasutas abivahendeid, märgendatud sõna, teksti tüüp, tema sotsiaalne taust, sugu, kodune keel ning haridus (Joonis 4). Et

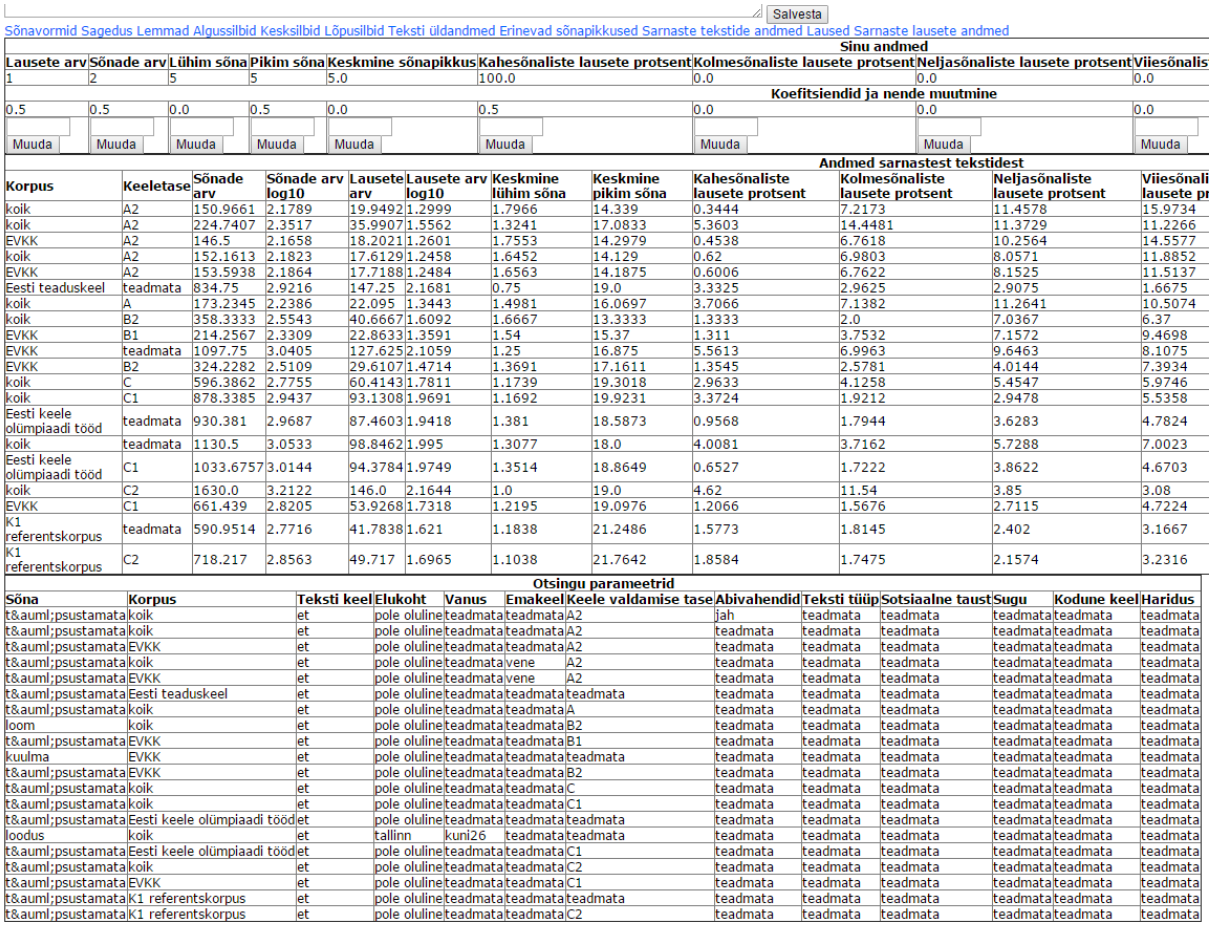
päring salvestuks dictionary-sse tuleb valida valikuks „statistika“.

Joonis 4. Päringu tegemine.

Kui vastav päring on tehtud, ilmub see silpide ja lemmade lehele, sarnaste lausete lehele otsingu parameetrite alla (Joonis 5). Tekst, millega võrdlust tehtud sai on varem hinnatud C keeletasemega essee.

```
Saage tuttavaks: normaalne inimene!


Tere!
Minu nimi on Eesnimi.
Mul on kaks kätt, kaks jalga, kaks silma, kaks kõrva, nina, suu.

Kas sellest piisab, et olla normaalne?
Visuaalne normaalsus annab kindlustunde ja rahulolu selleks, et tunnetada ja tunnistada end
normaalseks.
See on alusbaasiks mitmetele erinevatele komponentidele, mille alusel ühiskond ja teised selle
liikmed mind inimesena aktsepteerivad.

Olles harjumuspäraselt kuuluv kogukondatesse, milles erinevad etnilised, kultuurilised grupid,
sõpruskonnad on kehtestanud või kujundanud reeglid, peab inimene sotsialiseeruma.

Siin lisandub füüsilisele, kõige kergemalt, visuaalselt mõistetavasse vormi uus komponent.
Sotsialiseerumine algab hetkest, mil inimesehakatis puutub kokku, kommunikeerub maailmaga.
Iseõpitud ja õpitatavad sotsiaalsed normid kujunevad teatud aegruumis ja vastavas kontekstis,
milles inimene areneb.
Need kujunevad normideks või tabudeks.
Sotsialiseerumisprotsessi käigus formuleerub taluvuspiir, mis jääb kas ühele või teisele poole
normaalsust, milles kujunetakse.
See on norm.

Diskussioonis, teemal normist ja normaalsusest kumab läbi inimese olemuslik vajadus
kindlaksmääratud raamide ja reeglite järele.
Need justkui tagaksid eksistentsi kindluse.
Vaimne tasakaal sõltub suuremal määral traditsioonilisusel ja kindlustundel.
"Mõtlen, järelikult olen olemas".
Vaimne, emotsionaalne külg inimesest kujuneb ja areneb ennast teadvustades ja mõeldes.
On inimesi, kellele on oluline vaimse traditsiooni järgimine end teistega samastades.
Nii saadakse kinnitust oma minapildile.

Normaalsust inimese juures ei saa hinnata, seda saab tunnetada erinevate tahkudena, kuhu
kuuluvad füüsiline, vaimne ja sotsiaalne pool inimesest.
```

Tänapäeva multikultuurilises ja mitmekeelses maailmas on areng võimalik tänu erinevate normidele ja mitmetähenduslikkusele.
Iga inimene loob endast ise oma minapildi, mida aluseks võttes identifitseerib end normaalseks inimeseks, sest see peegeldab tagasi rühmast, millesse ta kuulub.[1]



Salvesta

Sõnavormid Sagedus Lemmad Algussilbid Kesksilbid Lõpusilbid Teksti üldandmed Erinevad sõnapikkused Sarnaste tekstide andmed Laused Sarnaste lausete andmed

Sinu andmed

Lausete arv	Sõnade arv	Lühim sõna	Pikim sõna	Keskmine sõnapikkus	Kahesõnaliste lausete protsent	Kolmesõnaliste lausete protsent	Neljasõnaliste lausete protsent	Viiesõnalis
1	2	5	5	5.0	100.0	0.0	0.0	0.0

Koefitsiendid ja nende muutmine

0.5	0.5	0.0	0.5	0.0	0.5	0.0	0.0	0.0
Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda

Andmed sarnastest tekstidest

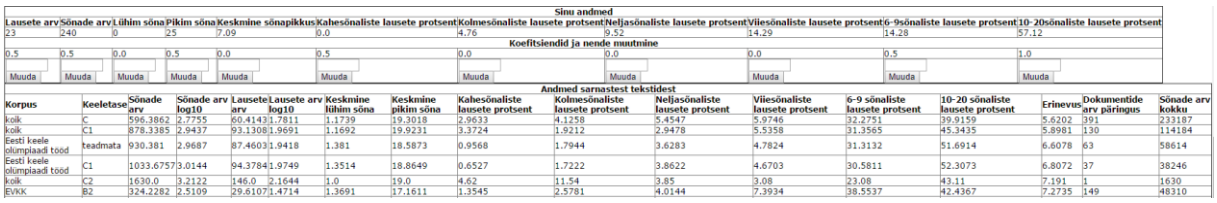
Korpus	Keeletase	Sõnade arv	Sõnade arv log10	Lausete arv	Lausete arv log10	Keskmine lühim sõna	Keskmine pikim sõna	Kahesõnaliste lausete protsent	Kolmesõnaliste lausete protsent	Neljasõnaliste lausete protsent	Viiesõnali lausete pr
koik	A2	150.9661	2.1789	19.9492	1.2999	1.7966	14.339	0.3444	7.2173	11.4578	15.9734
koik	A2	224.7407	2.3517	35.9907	1.5562	1.3241	17.0833	5.3603	14.4481	11.3729	11.2266
EVKK	A2	146.5	2.1658	18.2021	1.2601	1.7553	14.2979	0.4538	6.7618	10.2564	14.5577
koik	A2	152.1613	2.1823	17.6129	1.2458	1.6452	14.129	0.62	6.9803	8.0571	11.8852
EVKK	A2	153.5938	2.1864	17.7188	1.2484	1.6563	14.1875	0.6006	6.7622	8.1525	11.5137
Eesti teaduskeel	teadmata	834.75	2.9216	147.25	2.1681	0.75	19.0	3.3325	2.9625	2.9075	1.6675
koik	A	173.2345	2.2386	22.095	1.3443	1.4981	16.0697	3.7066	7.1382	11.2641	10.5074
koik	B2	358.3333	2.5543	40.6667	1.6092	1.6667	13.3333	1.3333	2.0	7.0367	6.37
EVKK	B1	214.2567	2.3309	22.8633	1.3591	1.54	15.37	1.311	3.7532	7.1572	9.4698
EVKK	teadmata	1097.75	3.0405	127.625	2.1059	1.25	16.875	5.5613	6.9963	9.6463	8.1075
EVKK	B2	324.2282	2.5109	29.6107	1.4714	1.3691	17.1611	1.3545	2.5781	4.0144	7.3934
koik	C	596.3862	2.7755	60.4143	1.7811	1.1739	19.3018	2.9633	4.1258	5.4547	5.9746
koik	C1	878.3385	2.9437	93.1308	1.9691	1.1692	19.9231	3.3724	1.9212	2.9478	5.5358
Eesti keele olümpiaadi tööd	teadmata	930.381	2.9687	87.4603	1.9418	1.381	18.5873	0.9568	1.7944	3.6283	4.7824
koik	teadmata	1130.5	3.0533	98.8462	1.995	1.3077	18.0	4.0081	3.7162	5.7288	7.0023
Eesti keele olümpiaadi tööd	C1	1033.6757	3.0144	94.3784	1.9749	1.3514	18.8649	0.6527	1.7222	3.8622	4.6703
koik	C2	1630.0	3.2122	146.0	2.1644	1.0	19.0	4.62	11.54	3.85	3.08
EVKK	C1	661.439	2.8205	53.9268	1.7318	1.2195	19.0976	1.2066	1.5676	2.7115	4.7224
K1 referentskorpus	teadmata	590.9514	2.7716	41.7838	1.621	1.1838	21.2486	1.5773	1.8145	2.402	3.1667
K1 referentskorpus	C2	718.217	2.8563	49.717	1.6965	1.1038	21.7642	1.8584	1.7475	2.1574	3.2316

Otsingu parameetrid

Sõna	Korpus	Teksti keel	Elukoht	Vanus	Emakeel	Keele valdamise tase	Abivahendid	Teksti tüüp	Sotsiaalne taust	Sugu	Kodune keel	Haridus
täpsustamata	koik	et	pole oluline	teadmata	teadmata	A2	jah	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	koik	et	pole oluline	teadmata	teadmata	A2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	EVKK	et	pole oluline	teadmata	teadmata	A2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	koik	et	pole oluline	teadmata	vene	A2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	EVKK	et	pole oluline	teadmata	vene	A2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	Eesti teaduskeel	et	pole oluline	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	koik	et	pole oluline	teadmata	teadmata	A	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
loom	koik	et	pole oluline	teadmata	teadmata	B2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	EVKK	et	pole oluline	teadmata	teadmata	B1	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
kuulma	EVKK	et	pole oluline	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	EVKK	et	pole oluline	teadmata	teadmata	B2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	koik	et	pole oluline	teadmata	teadmata	C	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	koik	et	pole oluline	teadmata	teadmata	C1	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	Eesti keele olümpiaadi tööd	et	pole oluline	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
loodus	koik	et	tallinn	kuni26	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	Eesti keele olümpiaadi tööd	et	pole oluline	teadmata	teadmata	C1	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	koik	et	pole oluline	teadmata	teadmata	C2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	EVKK	et	pole oluline	teadmata	teadmata	C1	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	K1 referentskorpus	et	pole oluline	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	K1 referentskorpus	et	pole oluline	teadmata	teadmata	C2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata

Joonis 5. Erinevad päringud.

Koefitsente muutes muutub ka erinevus, kuna tegemist on C keeletasemele vastava tekstiga, siis ka erinevus vastava C tasemega kõige väiksem (Joonis 6, joonis 7).Esialgsete koefitsentidega on erinevus kõige väiksem, varem hinnatud tekstide leitud informatsiooniga on praeguse teksti erinevus väiksem optimaalsete koefitsentidega.



Sinu andmed

Lausete arv	Sõnade arv	Lühim sõna	Pikim sõna	Keskmine sõnapikkus	Kahesõnaliste lausete protsent	Kolmesõnaliste lausete protsent	Neljasõnaliste lausete protsent	Viiesõnaliste lausete protsent	6-9sõnaliste lausete protsent	10-20sõnaliste lausete protsent
23	240	0	25	7.09	0.0	4.76	9.52	14.29	14.28	57.12

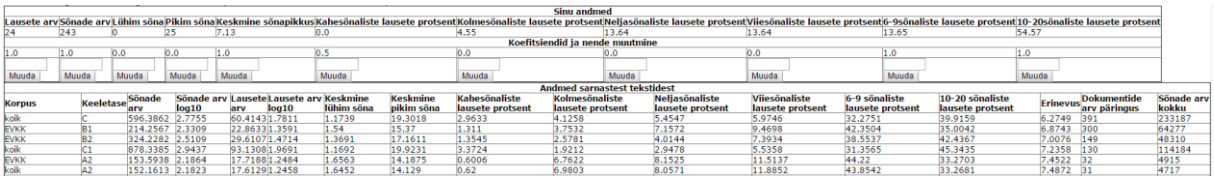
Koefitsiendid ja nende muutmine

0.5	0.5	0.0	0.5	0.0	0.5	0.0	0.0	0.0	0.5	1.0
Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda

Andmed sarnastest tekstidest

Korpus	Keeletase	Sõnade arv	Sõnade arv log10	Lausete arv	Lausete arv log10	Keskmine lühim sõna	Keskmine pikim sõna	Kahesõnaliste lausete protsent	Kolmesõnaliste lausete protsent	Neljasõnaliste lausete protsent	Viiesõnaliste lausete protsent	6-9 sõnaliste lausete protsent	10-20 sõnaliste lausete protsent	Erinevus	Dokumentide arv päringus	Sõnade arv kokku
koik	C	596.3862	2.7755	60.4143	1.7811	1.1739	19.3018	2.9633	4.1258	5.4547	5.9746	32.2751	39.9159	5.6202	391	233187
koik	C1	878.3385	2.9437	93.1308	1.9691	1.1692	19.9231	3.3724	1.9212	2.9478	5.5358	31.3565	45.3435	5.8981	130	114184
Eesti keele olümpiaadi tööd	teadmata	930.381	2.9687	87.4603	1.9418	1.381	18.5873	0.9568	1.7944	3.6283	4.7824	31.3132	51.6914	6.6078	63	58614
Eesti keele olümpiaadi tööd	C1	1033.6757	3.0144	94.3784	1.9749	1.3514	18.8649	0.6527	1.7222	3.8622	4.6703	30.5811	52.3073	6.8072	37	38246
koik	C2	1630.0	3.2122	146.0	2.1644	1.0	19.0	4.62	11.54	3.85	3.08	23.08	43.11	7.191	1	1630
EVKK	B2	324.2282	2.5109	29.6107	1.4714	1.3691	17.1611	1.3545	2.5781	4.0144	7.3934	38.5537	42.4367	7.2735	149	48310

Joonis 6. Esialgsed koefitsendid ja erinevus.



Sinu andmed

Lausete arv	Sõnade arv	Lühim sõna	Pikim sõna	Keskmine sõnapikkus	Kahesõnaliste lausete protsent	Kolmesõnaliste lausete protsent	Neljasõnaliste lausete protsent	Viiesõnaliste lausete protsent	6-9sõnaliste lausete protsent	10-20sõnaliste lausete protsent
24	243	0	25	7.13	0.0	4.55	13.64	13.64	13.65	54.57

Koefitsiendid ja nende muutmine

1.0	1.0	0.0	0.0	1.0	0.5	0.0	0.0	0.0	1.0	1.0
Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda	Muuda

Andmed sarnastest tekstidest

Korpus	Keeletase	Sõnade arv	Sõnade arv log10	Lausete arv	Lausete arv log10	Keskmine lihim sõna	Keskmine pikim sõna	Kahesõnaliste lausete protsent	Kolmesõnaliste lausete protsent	Neljasõnaliste lausete protsent	Viiesõnaliste lausete protsent	6-9 sõnaliste lausete protsent	10-20 sõnaliste lausete protsent	Erinevus	Dokumentide arv päringus	Sõnade arv kokku
koik	C	596.3862	2.7755	60.4143	1.7811	1.1739	19.3018	2.9633	4.1258	5.4547	5.9746	32.2751	39.9159	6.2749	391	233187
EVKK	B1	214.2567	2.3309	22.8633	1.3591	1.54	15.37	1.311	3.7532	7.1572	9.4698	42.3504	35.0042	6.8743	300	64277
EVKK	B2	324.2282	2.5109	29.6107	1.4714	1.3691	17.1611	1.3545	2.5781	4.0144	7.3934	38.5537	42.4367	7.0076	149	48310
koik	C1	878.3385	2.9437	93.1308	1.9691	1.1692	19.9231	3.3724	1.9212	2.9478	5.5358	31.3565	45.3435	7.2358	130	114184
EVKK	A2	153.5938	2.1864	17.7188	1.2484	1.6563	14.1875	0.6006	6.7622	8.1525	11.5137	44.22	33.2703	7.4522	32	4915
koik	A2	152.1613	2.1823	17.6129	1.2458	1.6452	14.129	0.62	6.9803	8.0571	11.8852	43.8542	33.2681	7.4872	31	4717

Joonis 7. Koefitsentide muutmine ja erinevuse muutumine.

Koefitsentide küsimine on seotud sessiooniga. Kui kasutaja on varem lehel käinud ja erinevaid muutatusi koefitsentidega teinud, siis kuvatakse järgmisel korral talle needsamad väärtused. Juhul kui pole muudatusi tehtud või esmakordselt määrtakse koefitsentidele tavapärased väärtused (Joonis 8).

```
def getKoef(self):
    "koefitsiendi kysimine"
    if self.REQUEST.SESSION.get('koef', 0):
       return self.REQUEST.SESSION.get('koef')
    else:
      koef= {
         'sonadearv': 0.5,
         'lausetearv': 0.5,
         'keskminesonapikkus': 0.0,
         'keskminelausepikkus':1.0,
         'keskminelyhimsona': 0.0,
         'keskminepikimsona': 0.5,
         '2wordpercent': 0.5,
         '3wordpercent': 0.0,
         '4wordpercent': 0.0,
         '5wordpercent': 0.0,
         '6to9wordpercent': 0.5,
         '10to20wordpercent': 1.0,
         '2sentencespercent': 0.5,
         '3sentencespercent': 0.0,
         '4sentencespercent': 0.0,
         '5sentencespercent': 0.0,
         '6to9sentencespercent': 0.5,
         '10to20sentencespercent': 1.0
      }
      self.REQUEST.SESSION.set('koef', koef)
      return koef
```

Joonis 8. Koefitsentide määramine.

Kasutaja ise saab koefitsente muuta kasutades vastavaid lahtreid ning ise trükkides või kasutades üles/alla nooli väärtuse seadmiseks, kui vastaval parameetril on soovitud väärtus, tuleb vajutada „Muuda“ nuppu. Seejärel muudetakse vastava parameetri koefitsent ning suunatakse kasutaja samale lehele tagasi (Joonis 9).

```
def setKoefParam3(self, REQUEST):
    "parameetri salvestus"
    k=self.getKoef()
    for param in k.keys():
      if REQUEST.get(param, 0):
        k[param]=float(REQUEST.get(param))
        #return "muudeti "+param
    REQUEST.RESPONSE.redirect('usertext.html?get2SentencesPercent')
```

Joonis 9. Tagasi suunamine.

3.3 Valmis tulemus

Tehtud töö on olemas Eesti vahekeele korpuse avalikul lehel, http://evkk.tlu.ee/wordtree/usertext.html ning http://evkk.tlu.ee/Search kui valida tulemusteks statistika. Samuti oli see aluseks ka Virgo Halliku tehtud seminaritööle "Eestikeeleste tekstide võrdluslehe täiendamine“.

Kokkuvõte

Eesmärk sai täidetud, funktsionaalsus sai loodud.

Kokkuvõtteks sai tehtud valmis suur lisa korpusele. Nüüd on võimalus korpuses kasutajal saada oluliselt rohkem informatsiooni oma teksti kohta.Töö venis esialgu, aga kui probleemidele lahendus leiti, sujus töö juba hästi. Töö leidis kasutust Virgo Halliku seminaritöös, kes tegi korpusel disaini moodsamaks.Töö käigus õppisin lähemalt tundma Pythoni programmeerimiskeelt. Samuti sain ka mõningase juhtimiskogemuse.

Kasutatud allikad

1. Eesti vahekeele korpus. (2016) http://evkk.tlu.ee/wwwdata/what_is_evk (07.03.2016)
2. Teadmata, teadmata, korpuse näidistekst esseena. http://evkk.tlu.ee/Documents/doc_740091150996_item?searchword=eesnimi
3. Python.(2016)Python https://www.python.org/ (07.03.2016)
4. Zope.(2016)Zope http://www.zope.org/ (07.03.2016)
5. Medijainen, S.2011. „Eesti vahekeele korpuse tekstide märgendusmooduli arendamine“. Haapsalu. Diplomitöö.
6. Hallik, V. 2015. „Eestikeelsete tekstide võrdlusleheküIje täiendamine“. Tallinn. Seminaritöö.

Lisad

Lisa 1

Seminaritööd puudutavad koodifailid asuvad aadressil www.tlu.ee/~hants/Seminaritoo.