Tallinna Ülikool
Digitehnoloogiate Instituut

Eestikeelsete tekstide võrdlusleheküle täiendamine

Seminaritöö

Autor: Virgo Hallik
Juhendaja: Jaagup Kippar

Autor:.....,,.....”2015
Juhendaja:.....,,.....”2015
Instituudi direktor:.....,,.....”2015

Tallinn 2015

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

Seminaritöö teema valikul oli määrav autori soov luua midagi praktilist ja kasulikku. Sellega seoses pakkus huvi eestikeelsete kirjalike tekstide võrdlemine, mis on üks etappidest teel eesti keele oskuse automaatse tuvastamiseni A2, B1, B2 ja C1-tasemetel.

Seminaritöö eesmärk on arendada Tallinna Ülikooli eesti vahekeele korpuse (edaspidi EVKK) tekstide analüüsi ja keeleoskustasemete võrdlemise lehekülge. Eesmärgi põhjendab vajadus mugavama ja samas avarama funktsionaalsusega lahenduse järele.

Seminaritöö ülesanneteks on 1. olemasoleva tekstide võrdlemise lehekülje eeliste ja puuduste väljatoomine ning parema lahenduse kavandamine; 2. DataTables'i tutvustamine ja põhjendus selle kasutamiseks andmete kuvamisel; alternatiivsete võimaluste tutvustamine; 3. arenduse kirjeldus ning töö tulemuste katsetamine.

Töötavat lahendust on võimalik vaadata aadressil
http://greeny.cs.tlu.ee:18187/korpus/korpus/wordtree/moodul.html?getText

1. Eesti vahekeele korpus

Selles peatükis tutvustatakse Eesti vahekeele korpust, selle juurde kuuluva tekstianalüüsi mooduli tekstide sarnasuste lehti ja tuuakse välja nende plussid ja miinused.

1.1. Eesti vahekeele korpuse tutvustus

Keeleteaduses mõeldakse sõna *korpus* all tavaliselt keeleainese kogu, mida kasutatakse uurimistöö materjalina. Tänapäeval mõeldakse korpuse all peamiselt polüfunktsionaalseid elektroonilisel kujul olevaid tekstikogusid, millesse kuuluvad tekstid on valitud eesmärgipäraselt, nii et nendest koosnev tervik annaks tõepärase pildi keelest (vt Muischnek, kuupäev puudub).

Eesti vahekeele korpus on eesti keele õppijate kirjalike tekstide kogu, mille praegune versioon on valminud Tallinna Ülikooli filoloogide, haridustehnoloogide ja informaatikute koostööna. Tegemist on monitorkorpusega, kuna sellesse lisatakse pidevalt uusi tekste.

Korpust saab kasutada

1) empiirilist ja rakenduslikku laadi uurimistöös,

2) tulevaste õpetajate ning lingvistide koolitamisel,

3) tegevõpetajate täiendõppes,

4) eesti keele õpetamisel ja individuaalõppes.

Korpuse funktsionaalsusi ja kirjalike tekstide analüüsi vahendeid saavad kasutada kõik, eriõigused on registreeritud kasutajatel, andmehalduril ja programmeerijal. (vt Eslon 2014, 438.)

1.2. Teksti analüüsi mooduli tekstide sarnasuse leht

Korpuse tekstianalüüsi moodulis on kaks lehekülge, et võrrelda enda sisestatud teksti andmeid tuumkorpuse A2, B1, B2 ja C1-taseme tekstide kohta tehtud päringute tulemustega. Esimene leht on mõeldud sõnade, teine lausete pikkuse võrdlemiseks keskmise näitaja alusel (vt Joonis 1), mõlemad lehed koosnevad neljast osast.

Esimeseks osaks on andmed sisestatud teksti kohta, sealhulgas vastavalt leheküljele sõnade või lausete pikkuse protsentidega.

Teises on esimeses osas välja toodud näitajatele koefitsientide seadmine, mida võetakse arvesse tekstide erinevuse arvutamisel.

Kolmas osa on varem tehtud päringute põhjal saadud tulemuste statistika, kus välja toodud päringule vastanud dokumentide keskmised andmed ja enda sisestatud teksti andmed – nende kahe võrdlemisel selgub erinevus nt A2, B1 ja B2-taseme näitajatest ning lähedus C1-taseme näitajatega jne. Erinevus on arvutatud sisestatud teksti andmete ning iga keeleoskustaseme tekstide kohta saadud andmete absoluutsete vahede korrutamisel neile vastavate koefitsientidega ning seejärel nende summa leidmisel.

Neljas ja viimane osa on tabel, kus näha varasemate päringute sooritamiseks kasutatud parameetrid.

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
1	2	5	5	5.0	100.0	0.0	0.0	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	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	A2	150.9661	2.1789	19.9492	1.2999	1.7966	14.339	0.3444	7.2173	11.4578	15.9734	48.1459	20.2702	13.2034	59	8907
koik	A2	224.7407	2.3517	35.9907	1.5562	1.3241	17.0833	5.3603	14.4481	11.3729	11.2266	38.4636	20.7459	13.4232	216	48544
EVKK	A2	145.9462	2.1642	18.172	1.2594	1.7527	14.2796	0.4587	6.8345	10.2472	14.7142	47.2534	24.0672	13.6454	93	13573
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	14.2599	31	4717
koik	A	173.2345	2.2386	22.095	1.3443	1.4981	16.0697	3.7066	7.1382	11.2641	10.5074	38.0607	25.2016	14.393	1305	226071
koik	B2	358.3333	2.5543	40.6667	1.6092	1.6667	13.3333	1.3333	2.0	7.0367	6.37	40.9633	38.82	14.4144	3	1075
EVKK	B1	214.4348	2.3313	22.8863	1.3596	1.5385	15.3813	1.3154	3.7658	7.1811	9.4569	42.3583	34.9429	15.7526	299	64116
EVKK	teadmata	1097.75	3.0405	127.625	2.1059	1.25	16.875	5.5613	6.9963	9.6463	8.1075	42.5375	15.9875	16.2743	8	8782
koik	C1	816.5376	2.912	92.6344	1.9668	1.0968	20.3441	4.4544	2.0003	2.584	5.8802	31.6651	42.5729	18.3141	93	75938
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	18.4351	149	48310
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	18.6793	391	233187
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	20.198	1	1630
EVKK	C1	661.439	2.8205	53.9268	1.7318	1.2195	19.0976	1.2066	1.5676	2.7115	4.7224	31.8217	52.2666	21.0924	41	27119

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	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	koik	et	pole oluline	teadmata	teadmata	C1	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	C2	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata
täpsustamata	EVKK	et	pole oluline	teadmata	teadmata	C1	teadmata	teadmata	teadmata	teadmata	teadmata	teadmata

Joonis 1. Vana tekstide sarnaste leht

Plussid

- Lehekülgede laadimiskiirus on hea.

Miinused

- Kasutatakse kahte lehekülge.
- Koefitsientide seadmine on tülikas.

- Puudub võimalus andmete sorteerimiseks.
- Puudub võimalus tabeli infot filtreerida.
- Andmeid saab võrrelda ainult varem tehtud päringutega.

2. Arenduse planeerimine

Selles peatükis on väljatoodud ideed mida üritatakse arenduse käigus realiseerida, kasutatavad tehnoloogiad, sealhulgas põhjendus Datatables'i kasutamiseks ja selle alternatiivid.

2.1. Arendusideed

Andmete filtreerimise ja sorteerimise võimalus

Hetkel on kuvatav info staatilises tabelis ning ei võimalda kasutajapoolset interaktsiooni, v.a koefitsientide muutmise. Filtreerimise ja sorteerimise võimaldamine loob eelduse vajaliku info kiiremaks leidmiseks, muutes tabeli kasutamise mugavamaks.

Enda sisestatud teksti võrdlemine teiste tekstidega

Olemasoleval lehel on võimalik enda sisestatud teksti võrrelda varem tehtud päringute keskmiste andmetega. Detailsema ülevaate annaks aga kindla eesmärgiga valitud tekstide andmete eraldi võrdlemine enda sisestatud teksti andmetega.

Lihtsam süsteem tekstide erinevuse arvutamiseks ja kontrolliks

Võrdluslehtedel on olemas küll koefitsientide muutmise võimalus, mis võimaldab määrata, kui palju mingisugust parameetrit tekstide erinevuse arvutamisel arvesse võetakse, kuid see on tavakasutamiseks liiga tülikas. Parem moodus oleks parameetrite lihtne sisse- ja väljalülitamine, tabelis kuvamise ja erinevusega arvestamise jaoks.

Kasutatavate korpuste valikuvõimalus

Olemasoleval lehel kuvatavate andmete jaoks on päringud varasemalt tehtud, lisatavate andmetabelite puhul oleks aga hea anda kasutajale võimalus valida, milliseid korpuseid kasutatakse.

2.2. Kasutatavad tehnoloogiad

Tekstide sarnasuse/erinevuse võrdlemise lehe arendustöös on kasutatud järgmisi tehnoloogiaid:

- Python - laialtlevinud kõrgema taseme programmeerimiskeel
- Zope – tasuta ning avatud lähtekoodiga objektorienteeritud veebirakenduse server, mis on kirjutatud Pythonis ning millele on ehitatud EVKK.
- JavaScript – tasuta ja avatud lähtekoodiga Netscape'i väljatöötatud skriptikeel, mis suudab suhelda HTML-keeles kirjutatud lähtekoodiga ja võimaldab muuta veebilehed interaktiivsemaks (vt Vallaste 2000)
- jQuery – kiire, väike ja võimalusterohke JavaScripti teek, mis muudab JavaScripti kirjutamise lihtsamaks

2.3. Datatables'i kasutamine tabelite kuvamisel

Enamus võrdlusleheküljel kuvatavast infost on parema loetavuse jaoks tabelitena, olemasolevas versioonis tehakse seda tavaliste HTML-tabelite kujul. HTML-tabel üksi ei anna aga eriti palju võimalusi kasutajapoolseks interaktsiooniks, näiteks info sorteerimiseks. Selleks, et tabeleid oleks mugavama kasutada ja need omaksid suuremat kasutajapoolset kontrolli kui lihtne andmete kuvamine, tuleb kasutada JavaScripti.

Seminaritöös võetakse tabelite visuaalse poole parandamiseks ja funktsionaalsuse tõstmiseks kasutusele Datatables. Põhjus Datatables'i kasutamiseks on autori varasem kokkupuude Datatables'iga ning asjaolu, et tegemist on hästi dokumenteeritud, pidevalt arendatava ja aktiivse kommuuniga pistikprogrammiga.

Datatables'i puhul on tegemist avatud lähtekoodiga Javascripti jQuery teegi pistikprogrammiga, mis lisab mistahes HTML-tabelile laiendatud interaktsiooni juhtelemendid. Datatables'i võimaluste alla kuuluvad näiteks paginatsioon, kiirotsing ja mitme tulba järgi järjestamine, dünaamiline tabelite loomine, tabelite oleku salvestamine ja palju muud. Lisafunktsionaalsusi aitavad tõsta ka mitmed laiendused, mida tuumprogrammis pole, kuid on võimalik hiljem juurde lisada. Datatables'i

kasutusele võtmiseks on mitu võimalust: 1) Datatables ja selle laiendused eraldi alla laadida või kasutada nende serveris olevaid faile; 2) luua vastavalt vajadustele pakk ja lisada enda sisestatud lehekülje päisesse valikute põhjal genereeritud lingid, sarnaselt eelnevaga on võimalik kasutada nii nende serveris olevat pakki kui ka laadida see alla ja paigutada omasse.

Datatables'i tuumprogrammi kasutamiseks on vajalik, et lehekülje päises oleks viited nii jQueryle kui ka Datatables'i Javascripti ja CSS failile (vt koodinäide 1).

```
<link rel="stylesheet" type="text/css" href="https://cdn.datatables.net/1.10.9/css/jquery.dataTables.min.css"/>
<script type="text/javascript" src="https://code.jquery.com/jquery-2.1.4.min.js"></script>
<script type="text/javascript" src="https://cdn.datatables.net/1.10.9/js/jquery.dataTables.min.js"></script>
```

Koodinäide 1. Datatables'i kasutamiseks vajalikud failid

Tabelisse on andmeid võimalik saada DOMist, Javascriptist ja Ajaxi abil JSONist ning serveritest. Peale eelpool toodud failide lisamist on Datatables'iga tavalise HTML-tabelile uute võimaluste lisamine lihtne. Vajalik on korrektsel kujul olevat HTML-tabel (vt koodinäide 2).

```
<table id="tabeli_ID" class="display" cellspacing="0" width="100%">
        <thead>
            <tr>
                <th>Nimi</th>
                <th>Vanus</th>
            </tr>
        </thead>
        <tbody>
            <tr>
                <td>Kristjan Kibuvits</td>
                <td>61</td>
            </tr>
            <tr>
                <td>Mari Maasikas</td>
                <td>35</td>
           </tr>
        </tbody>
 </table>
```

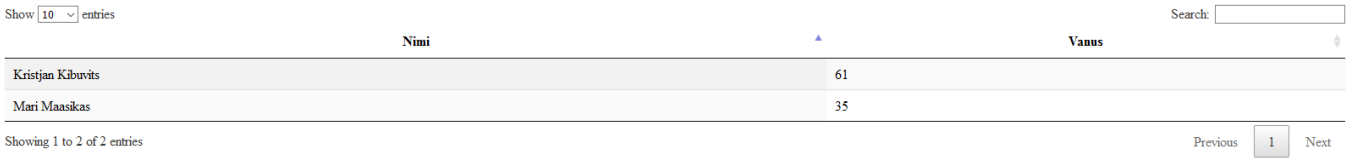
Koodinäide 2. Näidistabeli kood

Tabeli olemasolu järel lisada skript Datatables'i rakendamiseks (vt koodinäide 3).

```
$('#tabeli_ID').DataTable();
```

Koodinäide 3. Skript Datatables'i rakendamiseks

Kui kõik õnnestub, saab tabel uue kuju (vt joonis 2). Vaikimisi sättena on tabelitel olemas sorteerimine, otsing, leheküljed ja info.



Joonis 2. Näidistabel

Näitena toodud tabeli leiab aadressilt http://www.tlu.ee/~virgoh/tabelid/vtabel.html

Vaikimisi sätete paremaks nägemiseks ja katsetamiseks on olemas suurem näide aadressil

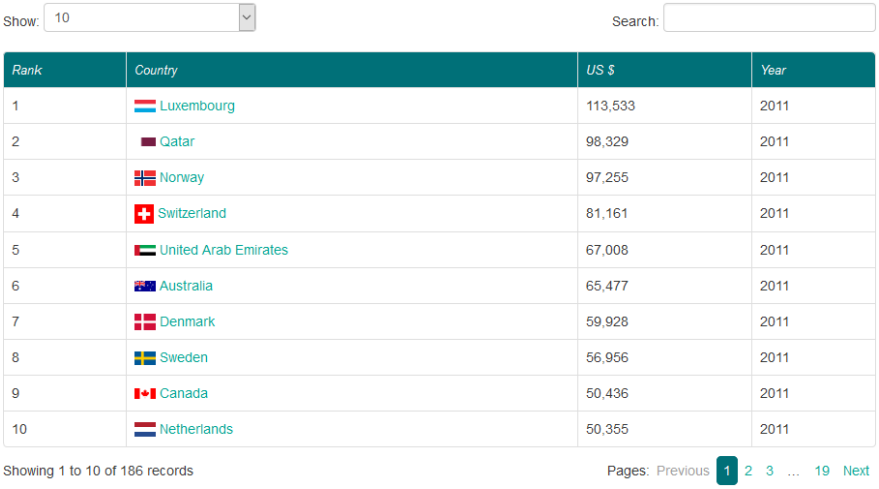
http://www.tlu.ee/~virgoh/tabelid/tabel.html

2.4. Datatables'i alternatiivid

On erinevaid lahendusi HTML-tabeli võimekuse tõstmiseks, kuid enamus neist piirdub sorteerimise lisamisega. Järgnevalt toob autor välja mõned Datatables'iga võimaluste poolest võrreldavad alternatiivid.

Dynatable

Dynatable on tasuta ja avatud lähtekoodiga jQuery pistikprogramm, mis kasutab tabelite (vt joonis 3) interaktiivseks muutmiseks järgmisi tehnoloogiaid: HTML5, JSON ja jQuery. Sarnaselt Datatables'ile on vaikimisi olemas sorteerimine, otsing ja paginatsioon.



Joonis 3. Näide Dynatable'i kodulehelt

Dynatable'i plussid

- Kiirus
- Paindlikkus
- Võimaldab piltide lisamist
- Sisseehitatud funktsioone on kasutajal lihtne täiendada

Datatables'i eelisteks Dynatable'i ees on veidi parem dokumenteeritus, suurem ja täienev lisafunktsionaalsuste hulk ja parem võimalus tekkinud küsimustele vastuste saamiseks.

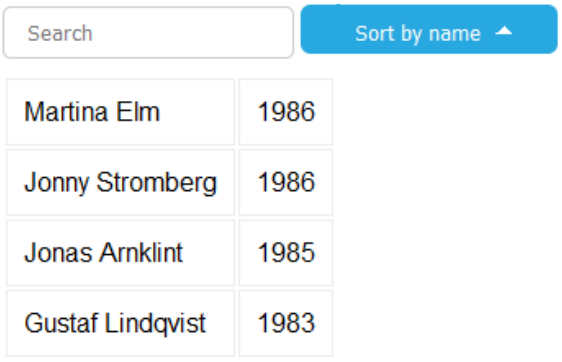
List.js

List.js on väike ja lihtne Javascripti teek mille abil on võimalik tabelitele (samuti nimekirjadele ja muudele erinevatele HTML-elementidele) lisada otsing, sorteerimine ja filtreerimine (vt joonis 4).

List.js plussid

- Lihtne paigaldada ja kasutada
- Kiire
- Väike
- Sõltuvusteta

List.js annab küll suurema vabaduse, kuid vajab tabelitele funktsionaalsuste lisamiseks veidike rohkem vaeva nägemist kui Datatables'i puhul.



Joonis 4. Näide List.js kodulehelt

3. Arendustöö

Käesolevas peatükis toon välja, millised muudatused sai tehtud olemasolevale tekstide võrdlemise osale, mida uut on lisatud ning milliseid funktsionaalsusi oli võimalik lisada Datatables'i ja selle laienduste abil.

EVKK arendustöid tehakse testkeskkonnas, et mitte häirida korpuse tavapärast kasutamist. Seminaritöö käigus tehtud arendustöö testkeskkonnaks on Tallinna Ülikooli Digitehnoloogiate instituudi testserver greeny.cs.tlu.ee, mille Zope serveris on juba varasemast ajast olemas EVKK testkeskkond ja millele ligipääsuks anti käesoleva arenduse autorile kõik õigused.

Tehtud töö võib jagada kaheks osaks: 1. muudatused ja edasiarendused, mis on tehtud funktsioonide failides, makrofailis ja lehekülje HTML poolel ning 2. lisafunktsionaalsused, mis on kuvatavatele tabelitele loodud Datatables'iga.

Arenduse käigus loodud ja muudetud failid leiab Lisa 1-st.

3.1. Tehtud muudatused ja edasiarendused

Lehtede ühendamine

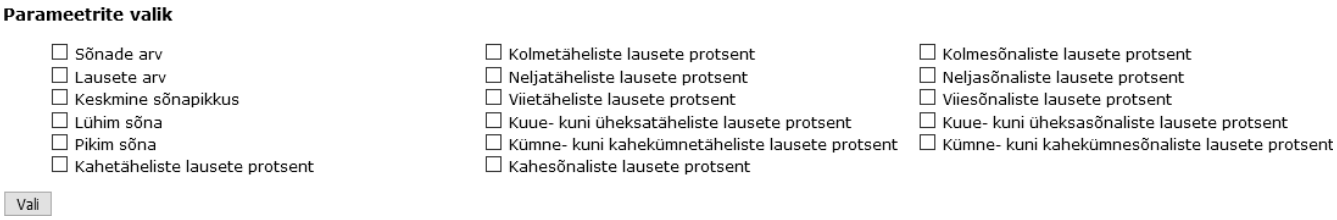
Olemasolevas versioonis tuli nii enda sisestatud teksti kui ka teiste tekstide kohta käivate sõnade ja lausete pikkuste osakaalu nägemiseks ja võrdlemiseks kasutada kahte eraldi lehekülge. Arenduse käigus tekkinud võimaluste tõttu ei olnud aga kahe lehekülje kasutamine vajalik ning tekstide sõna- ja lausepikkuste osakaalud kuvatakse samas tabelis.

Päringuinfo ja otsingu parameetrite tabelite ühendamine

Sarnaselt eelnevale ei ole tänu uutele võimalustele vajalik ka varasemalt tehtud päringute ja päringute parameetrite tabeleid eraldi kuvada. Mõlemas tabelis olnud andmed kuvatakse ühes tabelis.

Parameetrite valik

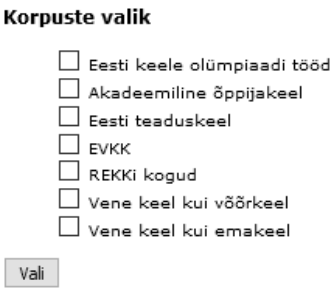
Koefitsientide seadmise süsteem annab küll kasutajatele parema kontrolli erinevuse arvutamise üle, kuid lisab kasutamisele liigset keerukust, mistõttu on see asendatud lihtsama märkeruutude süsteemiga(vt joonis 5). Parameetrite arvestamiseks erinevuse puhul ja nende kuvamiseks tabelites on vaja sobivate tulpade märkeruudud tähistada ja valik kinnitada. Tehtud valik salvestatakse sessioonis, seega pole korpuste valiku muutmisel vaja parameetreid uuesti valida.



Joonis 5. Parameetrite valik

Korpuste sisse- ja väljalülitamine

Uuele leheküljele on lisatud dokumentide päringu leheküljel kasutatav korpuste valiku makro. Selle abil on võimalik korpuse valikus (vt joonis 6) märkeruute tähistades valida, milliseid korpuseid kasutatakse uute andmetabelite juures päringute tegemiseks. Tehtud valik salvestatakse sessioonis, seega pole parameetrite valiku muutmisel vaja korpuseid uuesti valida. Esialgu on korpuse valik kasutatav ainult üksikute sarnaste tekstide tabeliga.



Joonis 6. Korpuste valik

Üksikute sarnaste tekstide tabel

Üksikute tekstide andmete võrdlemiseks enda sisestatud teksti andmetega on loodud uus tabel. Tabelis kuvatakse valitud korpustest pärit viiekümne teksti andmed, mille erinevus on kõige madalam. Kuvatavad andmed on vaikimisi järjestatud erinevuse põhjal kõige madalamast ehk

andmetelt sarnaseimast tekstist alustades. Erinevuse näitaja, sisestatud teksti ja korpustes olevate üksikute tekstide vahel saadakse parameetrite valikus märgitud andmete absoluutsete vahede summa leidmisel.

3.2. Datatables'i rakendamine

Uuel leheküljel kasutatakse DataTables'it enda sisestatud teksti andmete, üksikute tekstide andmete ja varasemate päringute andmete kuvamiseks ning interaktiivsemaks muutmiseks. Tänu uutele funktsionaalsustele muudab DataTables'i rakendamine lehe osade kasutamise mugavamaks. Lisatud on filtreerimine, sorteerimine, paginatsioon ja laienduse Buttons abil nupud tabelite vaadete muutmiseks. Järgnevalt on täpsemalt välja toodud, mis igale tabeli puhul on tehtud.

Enda sisestatud teksti andmete tabel

Enda sisestatud teksti andmete tabeli (vt joonis 7) puhul ei ole vaikimisi lisatavad funktsionaalsused vajalikud, küll aga kasutatakse tulpade vaikimisi peitmist ja lisatakse nupud kuva muutmiseks. Mittevajalikud funktsioonid nagu otsing, paginatsioon, info ja sorteerimine on võimalik välja lülitada, lisades need skripti väärtusega *false* (vt koodinäide 4).

```
"searching": false,
"paging": false,
"info": false,
"ordering": false
```

Koodinäide 4. Vaikimisi sätete väljalülitamine

Seejärel peidetakse vaikimisi vaatest sõnade ja lausete pikkuste osakaalude tulbad (vt koodinäide 5), eelnevalt on tulpadele lisatud klassid, vastavalt „sp” sõnapikkuste ja „lp” lausepikkuste puhul.

```
"columnDefs": [
        {
        "targets": ['sp','lp'],
         "visible": false
        }
              ]
```

Koodinäide 5. Tulpade peitmine vaikimisi vaates

Järgmisena on tabelile lisatud laienduse Buttons abil nupud, kõikide seminaritöös tehtud tabelinuppude jaoks on kasutatud nupu tüüpi *colVisGroup* mille abil on võimalik seada, millised tulbad on korraga nähtavad ja millised mitte. Nuppude loomiseks kasutatakse nuppude massiivi,

määratakse millist tüüpi nuppu kasutatakse, antakse nupule nimetus ja parameetrid. Selle tabeli jaoks on loodud kolm nuppu: “Näita sõnade protsente”(vt koodinäide 6), mille vajutades kuvatakse sõnade protsente, kuid peidetakse lausete protsendid, “Näita lausete protsente”, millele vajutades kuvatakse eelneva nupuga vastupidist ja “Taasta algne vaade”, mis taastab tabeli tavalise vaate.

```
{
   extend: 'colvisGroup',
   text: 'Näita sõnade protsente',
   show: '.sp',
   hide: '.lp'
}
```

Koodinäide 6. Nupud

Viimasena lisatakse skriptile DOM parameeter, milles seatakse, kuidas tabel ja selle juurde kuuluvad elemendid teineteise suhtes paiknevad. Selle tabeli puhul on see `dom:'Brt'` - nupud, töötlemine ja tabel.

Näita sõnade protsente | Näita lausete protsente | Taasta algne vaade

Enda teksti 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
19	286	2	26	7	8.33	0.0	0.0	8.33	50.0	41.66

Joonis 7. Enda sisestatud teksti andmete tabel

Varasemate päringute tabel

Varasemate päringute tabelis (vt joonis 8) on vaikimisi funktsioonidest välja lülitatud paginatsioon ja info. Vaikimisi vaatest on ka seekord peidetud osakaalude andmed. Sorteerimisele on lisatud väärtus tabeli vaikimisi sorteerimiseks: `"order": [[ 0, "asc" ]]` Nii on tabelis olevad read vaikimisi järjestatud esimese, st erinevuse tulba väärtuste järgi madalaimast alustades. Lisatud on kolm nuppu, mis olemas eelneval tabelil ja lisaks nupp „Näita otsingu parameetreid”, mis vahetab tabeli vaate päringute parameetrite peale. Tabeli jaluses on võimalik tabelit korpuse ja keeleoskustaseme (A2, B1, B2, C1) järgi filtreerida. Tabelile on lisatud horisontaalne kerimine, juhuks kui tabel täielikult arvutiekraanile ära ei mahu. Sarnaselt eelnevale on tabeli DOM parameetri väärtus „Brt”.

Näita sõnade protsente | Näita lausete protsente | Näita otsingu parameetreid | Taasta algne vaade

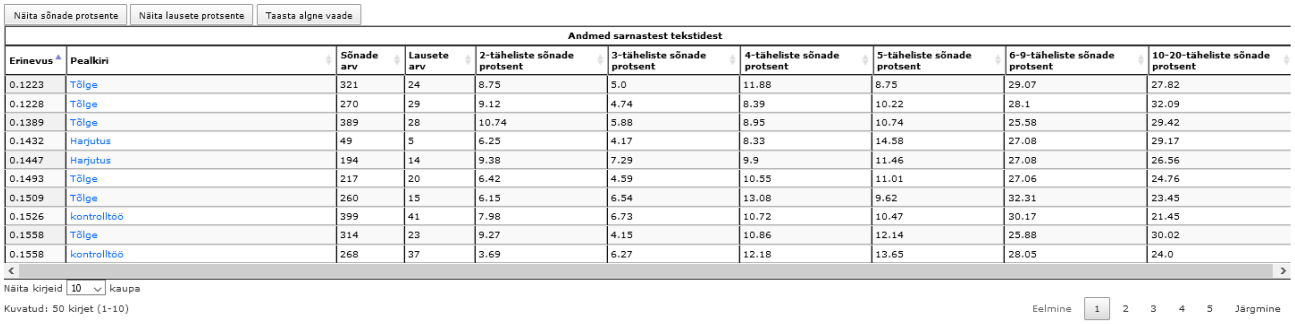
Andmed viimastest päringutest											
Erinevus	Korpus	Keeletase	Sõnade arv	Sõnade arv log10	Lausete arv	Lausete arv log10	Keskmine lühim sõna	Keskmine pikim sõna	Keskmine sõnapikkus	Dokumentide arv päringus	Sõnade arv kokku
9.2894	Akadeemiline õppijakeel	teadmata	1441.4074	3.1588	242.6296	2.3849	1.037	24.4074	6.4742	27	38918
14.1347	EVKK	C1	661.439	2.8205	53.9268	1.7318	1.2195	19.0976	5.8873	41	27119
15.4963	Eesti keele olümpiaadi tööd	teadmata	930.381	2.9687	87.4603	1.9418	1.381	18.5873	5.3709	63	58614
15.6163	EVKK	B2	324.2282	2.5109	29.6107	1.4714	1.3691	17.1611	5.5785	149	48310
16.7238	EVKK	teadmata	278.5907	2.445	28.228	1.4507	1.6244	16.1166	5.3578	386	107536
17.4025	EVKK	B1	214.4348	2.3313	22.8863	1.3596	1.5385	15.3813	5.2607	299	64116
17.417	koik	teadmata	259.4998	2.4141	34.2387	1.5345	1.256	16.0461	5.3719	11665	3027065
18.4672	EVKK	A2	145.9462	2.1642	18.172	1.2594	1.7527	14.2796	5.0474	93	13573
20.0714	koik	B2	358.3333	2.5543	40.6667	1.6092	1.6667	13.3333	4.9895	3	1075

Erinevus			Sõnade arv	Sõnade arv log10	Lausete arv	Lausete arv log10	Keskmine lühim sõna	Keskmine pikim sõna	Keskmine sõnapikkus	Dokumentide arv päringus	Sõnade arv kokku

Joonis 8. Varasemate päringute andmete tabel

Üksikute tekstide tabel

Üksikute tekstide tabeli (vt joonis 9) puhul lülitatakse vaikimisi seadetest välja ainult otsing. Erinevalt eelmistest tabelitest on kasutusel leheküljed ja info, ühel leheküljel kuvatavate ridade arvu on võimalik muuta ning info puhul on kasutatud eestikeelset tõlget. Sarnaselt eelnevale tabelile on vaikimisi sorteerimine erinevuse järgi, lisatud on horisontaalne kerimine, vaikimisi vaatest on peidetud osakaalud. Tabeli juures kasutatakse samu nuppe, mis enda sisestatud teksti andmete tabeli puhul. Tabeli DOM väärtus on seekord „Brtlip” - nupud, töötlemine, tabel, keel, info ja paginatsioon.


Näita sõnade protsente | Näita lausete protsente | Taasta algne vaade

Andmed sarnastest tekstidest

Erinevus	Pealkiri	Sõnade arv	Lausete arv	2-täheliste sõnade protsent	3-täheliste sõnade protsent	4-täheliste sõnade protsent	5-täheliste sõnade protsent	6-9-täheliste sõnade protsent	10-20-täheliste sõnade protsent
0.1223	Tõlge	321	24	8.75	5.0	11.88	8.75	29.07	27.82
0.1228	Tõlge	270	29	9.12	4.74	8.39	10.22	28.1	32.09
0.1389	Tõlge	389	28	10.74	5.88	8.95	10.74	25.58	29.42
0.1432	Harjutus	49	5	6.25	4.17	8.33	14.58	27.08	29.17
0.1447	Harjutus	194	14	9.38	7.29	9.9	11.46	27.08	26.56
0.1493	Tõlge	217	20	6.42	4.59	10.55	11.01	27.06	24.76
0.1509	Tõlge	260	15	6.15	6.54	13.08	9.62	32.31	23.45
0.1526	kontrolltöö	399	41	7.98	6.73	10.72	10.47	30.17	21.45
0.1558	Tõlge	314	23	9.27	4.15	10.86	12.14	25.88	30.02
0.1558	kontrolltöö	268	37	3.69	6.27	12.18	13.65	28.05	24.0

Näita kirjeid 10 kaupa

Kuvatud: 50 kirjet (1-10)

Eelmine 1 2 3 4 5 Järgmine

Joonis 9. Üksikute tekstide tabel

Kokkuvõte

Seminaritöö peamised eesmärgid ja ülesanded said täidetud. Kõikide andmete kuvamine on viidud ühele leheküljele muutmata tabelite kasutamist halvemaks. Üksikute tekstide andmete võrdlemiseks on lehele lisatud uus tabel. Datatables'i rakendamine andis kasutajale suurema kontrolli tabelite üle ja muutis lehe kasutamise mugavamaks.

Loodud lahendus andis avaramad võimalused tekstide andmete võrdlemiseks, tõstes lehe kasutamise mugavust lehte keerulisemaks muutmata.

Töö käigus sain uusi kogemusi Pythoni ja Zope kasutamisel, tutvusin põhjalikumalt pistikprogrammiga Datatables ning sain kogemuse selle oskuslikumaks kasutamiseks.

Kasutatud allikad

Eslon, P. (2014) Eesti vahekeele korpus. Keel ja kirjandus. 6, 436-451. http://kjk.eki.ee/ee/issues/2014/6/507 (18.10.2015).

Muischnek K. (Kuupäev puudub). Korpuslingvistika kursus:1 http://www.cl.ut.ee/kursused/korp_ling01(18.10.2015).

Vallaste, H. (2015). e-Teatmik. http://www.vallaste.ee/ (18.10.2015).

Zope. (2015).Zope http://www.zope.org/ (18.10.2015).

Python. (2015) Python https://www.python.org/ (18.10.2015).

jQuery. (2015). jQuery. http://jquery.com/ (18.10.2015).

Datatables. (2015). Datatables http://www.datatables.net/ (18.10.2015).

Dynatable. (2015). Dynatable http://www.dynatable.com/ (18.10.2015).

List.js.(2015).List.js. http://www.listjs.com (18.10.2015).

Lisad

Lisa 1

Seminaritöö käigus loodud ja muudetud lähtekoodi failid asuvad tööga kaasas oleval CD-plaadil ja veebis aadressil www.tlu.ee/~virgoh/Seminaritoo

Failide loetelu:

Document.py - Dokumendi funktsioonide fail. Lisatud funktsioonid andmete mugavamaks kättesaamiseks.

WordTree.py - Tekstianalüüsi funktsioonide fail. Lisatud üksikute tekstide erinevuse leidmise funktsioon, muudetud, muudetud varasemate päringute erinevuse leidmise funktsiooni ja parameetrite salvestamine.

macros.pt – Lisatud uus korpuste valimise makro ja parameetrite valimise makro.

moodul.pt – Uus tekstide võrdlemise leht.