

Kollokatsioonide tuvastaja kasutusjuhend

Kollokatsioonide tuvastaja koosneb kahest alamprogrammist: *SurfaceCooccurrence.pl* ja *CollocationFinder.jar*.

SurfaceCooccurrence.pl - võtab sisendiks korpusefailide kausta, kus asub vähemalt üks tarkvaraga t3mesta morfoloogiliselt ühestatud korpusefail laiendiga .t3 ja salvestab failide põhjal andmebaasi vajalikud sõnad ja sagedused.

CollocationFinder.jar - ühendub andmebaasi ja kasutaja poolt sisestatud sõna põhjal leiab sealt antud sõna kõik võimalikud kollokaadid koos seose tugevuse määraga, mis iseloomustab sõnadevahelist tõmbumist.

KOLLOKATSIOONIDE TUVASTAJA KASUTAMISE EELTINGIMUSED

- Masinasse on paigaldatud MySQL andmebaasi server ja on loodud andmebaas kodeeringuga *utf-8*.

MySQL tarkvara (soovitavalt viimase versiooni) saab alla laadida siit:
<http://www.mysql.com/downloads>

MySQLi installeerimisel tekitatakse automaatselt *root* kasutaja. Kui soovitakse luua mingi muu kasutaja, saab seda teha (esalt küll *root* kasutajaga sisse logides) käskudega:

```
create user '[uus kasutajanimi]'@'%' identified by '[password]';  
grant all privileges on *.* to '[uus kasutajanimi]'@'%' with grant option;  
(nurksulgudes olev (nurksulud k.a) asendada tegeliku soovitud väärtusega).
```

Sobiva kodeeringuga (UTF-8) andmebaasi saab luua käsuga:

```
create database [andmebaasi nimi] character set = utf8;
```

või eraldi käskudega:

```
andmebaasi loomine: create database [andmebaasi nimi];
```

```
andmebaasi kodeeringu muutmine: alter database [andmebaasi nimi] charset=utf8;
```

UNIX – avada terminali aknas MySQL käsuri, sisestada käsk:

```
mysql -u [kasutaja] -h [andmebaasiserver] -p
```

sisestada parool ja seejärel ülevalpool toodud käsk (või käsud) andmebaasi loomise kohta

WINDOWS - avada MySQL käsuri, st *MySQL Command Line Client*, sisestada kõigepealt parool (kui kasuviip näitab, et sisestada kasutaja, siis sisestada esalt kasutajanimi ja seejärel parool) ning siis käsud andmebaasi loomise kohta

- Masinasse on paigaldatud Perl koos vajalike teekidega DBD::mysql ja DBI.

UNIX - enamasti on Perl juba installeeritud. Versiooni saab kontrollida käsuga "perl -v". Vajadusel saab alla laadida siit: <http://www.activestate.com/activeperl/downloads>

WINDOWS - Perli (soovitavalt värskema versiooni) saab alla laadida siit: <http://strawberryperl.com>

Uuemate versioonidega on eelpool nimetatud teegid juba kaasas. Kui ei ole (ja programm annab käivitades vastava vea), siis saab neid paigaldada Perli teekide veebilehelt <http://www.cpan.org/>

UNIX – esmalt: *perl -MCPAN -e shell*
 et siseneda CPAN terminali ja seejärel: *install [moodul]::[nimi]*
 Vt täpsemalt: <http://perl.about.com/od/packagesmodules/qt/perlcpn.htm>
 WINDOWS - käsuga: *cpan [moodul]::[nimi]*

DBI – (Vt täpsemalt: <http://search.cpan.org/~timb/DBI-1.622/DBI.pm>)

DBD::mysql - (Vt: <http://search.cpan.org/~capttofu/DBD-mysql-4.022/lib/DBD/mysql.pm>)

- Masinasse on paigaldatud Java Runtime Environment (versioon 6 või 7).

Vajadusel installeerida nt siit: <http://www.java.com/en/download/index.jsp>

- Masinas on olemas kaust, mis sisaldab vähemalt üht korpusefaili laiendiga .t3

.t3 korpusefail saadakse morfoloogilise ühestaja t3mesta kasutamise tulemusel ja seal esinevad märgendid <s> ja </s>, mis tähistavad lause algust ja lõppu, ning <ignoreeri> ja </ignoreeri>, mille vahele jääb informatsioon lause autori, allika jms kohta. Lause moodustavad sõnad asuvad igaüks eraldi real. Igale sõnale järgneb vähemalt üks tema võimalik morfoloogiline analüüs.

Programmi loomisel ja testimisel kasutati selliseid tarkvara versioone:

- Windows Vista Home Premium masinas:
 - MySQL Server 5.5,
 - Strawberry Perl 5.16.2.1,
 - JDK 1.7.0_09 (kompileeritud siiski vastavuses versiooniga 1.6)
 - Java Runtime Environment 7 (1.7.0)
- Linux CentOS 5.8 masinas (ats.cs.ut.ee serveris):
 - MySQL Server 5.0.95
 - Perl 5.8.8
 - Java Runtime Environment 6 (1.6.0)

KOLLOKATSIOONIDE TUVASTAJA KÄIVITAMINE

I *SurfaceCooccurrence.pl*

1) Avada *SurfaceCooccurrence.pl* fail tekstiredaktoriga ja muuta sobivaks järgmised parameetrid, (jutumärkide vahele kirjutada õige väärtus vastavalt oma MySQL seadetele):

- andmebaasi server: `$host = ""`;
juhul kui server asub samas arvutis, kui kollokatsioonide tuvastaja: `$host = "localhost"`;
muul juhul serveri nimi, nt `$host = "ats.cs.ut.ee"`;
- andmebaasi nimi: `$db = ""`;
- MySQL kasutajatunnus: `$user = ""`;
- kasutaja parool: `$password = ""`;
- korpusefailide asukoht arvutis, st kausta nimi koos täisteedega: `$corpus_folder = ""`;
(jälgida, et kaustanime lõppu jääks kaldkriips / ja sisestatu ei sisaldaks täpikähti), nt
`$corpus_folder = "C:/korpus/"`;

2) Käivitada programm.

UNIX/WINDOWS – käsurealt (Windowsi puhul Perli käsurealt) liikuda kausta, kus asub programmfail: `cd [SurfaceCooccurrence.pl asukoht]`

ja siis käivitada programm: `perl SurfaceCooccurrence.pl`

Programm käivitub ja on töötamise lõpetanud, siis kui järgmine käsurea viip ilmub ekraanile.

NB! Kui failide lugemine ebaõnnestus (nt täpikähtede tõttu faili nimes või mõne kausta nimes) ja näidatakse viga „*error in opening dir*“, siis tuleb enne uut katset andmebaasi juba tekkinud tabelid ära kustutada. Selleks avada MySQL käsurida (vt ülalt) ja sisestada käsud:

```
use [andmebaasi nimi];  
drop table ff;  
drop table fl;  
drop table lf;  
drop table ll;  
drop table word;  
show tables;
```

Kui pärast viimase käsu sisestamist näidatakse „empty set“ on kõik tabelid andmebaasist edukalt kustutatud ja võib uuesti proovida *SurfaceCooccurrence.pl* käivitamist (korrektse korpuse failide kataloogide asukohaga, vt ülalt)

II *CollocationFinder.jar*

UNIX/WINDOWS - käsurealt (Windowsi puhul `run > cmd`):

```
java -jar [tee failini]/CollocationFinder.jar
```

või topletklõps failil *CollocationFinder.jar*

1) Sisestada programmi ülemisse vasakusse nurka andmebaasi parameetrid (host, andmebaasi nimi, kasutaja, parool), vastavalt sellele, mis sai muudetud programmfailis *SurfaceCooccurrence.pl*

2) Valida otsingusõna ja otsitavate kollokaatide kuju vastavalt soovile:

Otsingusõna:

sõnavorm - sisestatud otsingusõna on mingi sõnavorm, nt *tegi* või *lamba* jne

lemma - sisestatud otsingusõna on lemma, nt *tegema*, *lammas* jne

Kollokaadid:

sõnavorm - sisestatud sõna kollokaatidena vaadeldakse kõiki sõnavorme, nt *tegi*, *teeb* ja *tegema* loetakse erinevateks kollokaatideks

lemma - kollokaatidena vaadeldakse lemmasid, nt kui otsisõna esineb nii *tegi*, *teeb*, kui *tegema* ümbruses, siis loetakse kollokaadiks lemma *tegema*

3) Sisestada otsingusõna vastavasse lahtrisse.

4) Vajutada nupule "*Otsi kollokaate*"

Programmiakna paremasse ossa kuvatakse sisestatud sõna kõik võimalikud kollokaadid koos vastavate seoste tugevustega või teade, et kollokaate ei leitud, juhul kui sisestatud sõna andmebaasis ei esine või juhul kui see küll esineb, aga mitte ühegi teise sõnaga ühes aknas.

NB! Kui käivitamisel tekib probleeme sisestatud käsu ära tundmisega, siis kontrollida, et PATH keskkonna muutujas on olemas tee Perli asukohale arvutis. Võimalik, et sinna on vaja lisaks veel teed Perli teistele *bin* kaustadele.

UNIX - olemasolevaid keskkonnamuutujaid näeb käsuga *set* Perli asukoha saab PATH keskkonnamuutujale lisada käsuga: *export PATH=\${PATH}:[tee Perli bin kataloogi]*

WINDOWS - versiooniti pisut erinev. Aga üldjoontes: *Control Panel > System > (Advanced System Settings) > Environment variables > valida PATH > Edit > lisada [Perli kataloog]/c/bin/; [Perli kataloog]/perl/bin/; [Perli kataloog]/perl/site/bin;*
või *Commad Prompt* käsurealt:

echo %PATH%, et näha, mis on juba PATH muutujasse lisatud

kui eelmise käsu tulemusel kuvatud PATH muutuja väärtuste hulgas ei ole perli katalooge, sisestada *set PATH=[Perli bin kataloog]* (säilib ainult sessiooni lõpuni) või *setx PATH [Perli bin kataloog]*