Kasutusjuhend

Vajaminevad programmid

- *Perl* 5.14 või uuem
- *Lingua::Ident*
- Linux

Perl'i saab alla laadida leheküljelt http://www.perl.org/get.html .

Kui *Perl* on installitud, siis soovituslikult administraatori õigustes *Lingua::Ident*'i installimiseks tuleb konsoolis käivitada:

perl -MCPAN -e shell
cpan > **install Lingua::Ident**
cpan > **exit**

Programmi kasutamine

perl tuvasta_keel.pl [**-utf8**]

-utf8 lipu kasutamine on valikuline, märkimisel kasutatakse programmmi sissekirjutatud UTF-8 kodeeringus treeningfaile ja töös muudetud *Ident.pm* faili(*Ident_muudetud.pm*). Algselt on programmi sisse kirjutatud eesti, inglise ja saksa keele treeningfailid nii UTF-8 kodeeringus kui ilma.

Programm ootab sisendit ning tagastab eesti keele puhul sisendi enda, võõrkeele puhul märgendab selle <foreign> *tag*'idega.

Näiteks käsk

echo „On see lause eesti keeles“ | perl tuvasta_keel.pl

tagastab

on see lause eesti keeles
Lauseid oli failis 1

ja mitte-eestikeelsete lausete puhul nt.

echo „some foreign language“ | perl tuvasta_keel.pl

tagastab

<foreign keel='mitte_eesti'>some foreign language</foreign>
Lauseid oli failis 1

.

Vajadusel saab programmis teisi treeningfaile kasutada. Selleks tuleb teha treeningfailid mõnest sisendfailist. Lisadena kaasas olevate XML failide treeningul kasutamisel tuleb paremate tulemuste

saamiseks nendest ebavajalikud sümbolid eemaldada.

Kui UTF-8 kodeeringus treeningfaile vaja pole siis toimub treeningfailide tegemine algse **trainlid** utiliidi abil:

trainlid *sisendfaili keel < sisendfail > väljundfail*

näiteks

trainlid eesti < eesti_lyhike_naide.txt > eestinaide.trainlid

(Selline treeningfail on ainult näide, ei tuvasta hästi kuna sisaldab ainult ühte lauset)

Kui on vaja UTF-8 kodeeringus treeningfaile tekitada siis tuleb kasutada lõputöö käigus muudetud **trainlid** utiliiti:

perl trainlid.pl *sisendfaili keel < sisendfail > väljundfail*

näiteks

perl trainlid.pl eesti < eesti_lyhike_naide.txt > eestinaide.trainlid

(Selline treeningfail on ainult näide, ei tuvasta hästi kuna sisaldab ainult ühte lauset)

Kui uued treeningfailid on olemas tuleb nende kasutamiseks koodis vastavalt kodeeringule failinimed muuta või lisada (UTF-8 puhul reale 15, muidu reale 18 failis *tuvasta_keel.pl)*.

Kui esineb selline viga või analoogne

utf8 "\xB6" does not map to Unicode at .... line 6 3, <MATRIX> line 3974.

siis on põhjus tõenäoliselt selles, et treeningfail pole UTF-8 kodeeringus, aga kasutatakse muudetud *Ident.pm* faili (*Ident_muudetud.pm*), mis seda ootab.

Delfi failide automaatseks tuvastamiseks on olemas **tuvastafailid.sh** , mille käivitamisel töödeldakse kaustas *kommentaarid* asuvad delfi kommentaarid, tuvastatud keel (eesti või mitte-eesti) ja tulemused kirjutatakse kausta tulemused ja vastavalt kas *delfi1-9* või *delfi1-9_utf* faili. Lisaks kasutatab **tuvastafailid.sh** utiliiti *diff* et leida UTF-8 ja algse programmiga tekitatud erinevused ning salvestab need faili *erinevused.diff*.

Näide tuvastafailid.sh kasutamisest.

```
$ ./tuvastafailid.sh
Töötlen: kommentaarid/delfi1.xml
Lauseid oli failis 3655
Töötlen: kommentaarid/delfi1.xml_utf
Lauseid oli failis 3655
Töötlen: kommentaarid/delfi2.xml
Lauseid oli failis 880
Töötlen: kommentaarid/delfi2.xml_utf
Lauseid oli failis 880
...
```