

Reeglipõhine keeletarkvara

Jan Willemson, Jaak Pruulmann-Vengerfeldt
EKKTT konverents
Tallinn, 20.11.2007

Plaan

- Taust (Kust?)
- Eesmärgid (Kuhu?)
- Meetodid (Kuidas?)
- Tulemused (Kui kaugel?)
- EKKTT (Kes ja mille eest?)

Taust

- Tegelikult oleme me tavakasutajad..
vabavaraliste nišisüsteemide(?)
tavakasutajad
- Spelleripuudus'98. Filosoofil oli speller küll,
aga enamasti polnud sellega midagi teha
 - ainult valitud platvormidele, intellektuaalne omand
 - TeX, LyX, vabavaralised redaktorid – de facto spelleriliides: ispell
 - esimene tõsine katse: Pearu Peterson 2000 (ÕS 76 baasil)
 - Jan (aga mitte ainult?) otsis tudengit

Taust (2)

- Kevadel 2003 sai valmis Ülle Viksi “Väikesel Vormisõnastikul” baseeruv ispell'i sõnastik
 - ~15 miljonit sõnavormi (~282000 „lemmat“, ~9000 reeglit 26s grupis)
 - liitsõnad “kõik sobib kõigega”
 - palju “tavalisi” sõnu puudu
 - see on kõigest staatiline sõnastik

Taust (3)

- Mõistlik eesti keele speller on morfoanalüsaator
 - Filosofti ESTMORF on sisuliselt kommertsprogramm
 - EKIs VVS baasil tehtu vajab uuendamist ja pole praegusel kujul otse kasutatav
 - EKI reeglikomplekte peaks ka (edasi)arendama

Eesmärgid

- Luua EKIs loodud reeglipõhise eesti keele (morfoloogia) kirjelduse jaoks värskem rakendamismootor
 - mootor ja reeglid lahku, teisendusvahendid
- Täiendada reegleid, kus see on vajalik ja võimalik
 - liitsõnad, tuletised, testimine korpustega (avatud morfoloogiamudel), metainfo?
- Ehitada reeglite baasil analüüsiv speller
- Teha seda kõike nii, et tekkiv teadmus oleks lihtsalt kasutatav ka teistele
 - EKKTT sees ja sellest väljas

Kuidas?

- Olemasoleva EKI reeglipõhise tarkvara ja reeglite analüüs, sh:
 - EKI programmide taasteostamine uuemate vahenditega
 - mugav(am)ate ja kiire(ma)te vahendite loomine reeglite uuendamiseks
- Teisendusvahendid
 - reeglite ja tugisõnastike automaatsed teisendused teistesse esitustesse
 - konkreetsete rakenduste “kompileerimine”. Alustades spellerist

Kus me oleme?

- Osa EKI reeglitest (tüveteisendused, mõnede nimisõnatüüpide moodustamine) on taasteostatud
 - vahendid: Haskell, Functional Morphology (Markus Forsberg, Aarne Ranta)
 - süntees on lihtsam kui analüüs..
- Palju aega on kulunud olemasolevate muude asjadega tutvumisele

EKKTT

- Inimesed:
 - Jaak Pruulmann-Vengerfeldt
 - Jan Willemson
 - Ülle Viks
 - Indrek Hein
- Rahad:
 - 290 000 kr aastal 2007
- Mured:
 - litsentsid.. ?
- Tulevik?

Aitäh