Keeletehnoloogiast ja reeglipõhistest meetoditest

Sissejuhatus informaatikasse,

2. detsember 2010

Krista Liin

Keeletehnoloogia

- Mis on keeletehnoloogia ja milleks seda vaja on
- Reeglipõhised meetodid (vs muud võimalused)
- Töövõimalused
- Millega üks keeletehnoloog tegeleb

Keeletehnoloogia

Loomuliku keele töötlus arvutil

- Speller
- Otsing: sõnad, sõnavormid, sõnaliigid, tähenduskategooriad
- Spämmifilter
- Meediamonitooring: kes mida meist räägib?
- Kontrollitud keel ja automaatne küsimustele vastamine
- ...

Keeletehnoloogia

- Kõnetuvastus ja –süntees
- Sisukokkuvõtted veebilehtedest
- Mobiilid: internet, hääljuhtimine, kõnekeskused
- Masintõlge
- Dialoog loomulikus keeles
- Semantika
- ...

Keeletehnoloogia: meetodeid

- Lingvistiline teadmus
  - Rääkida oskan -> oskan ka arvutile selgeks teha
  - Vaja läheb: inimest ja keeleteooriaid, hindamist
- Korpuslingvistika
  - Vaatame olemasolevaid tekste, õpetame arvutit selle põhjal
  - Vaja läheb: inimest, keelelist teadmist ja suuri korpusi, hindamist

Keeletehnoloogia: meetodeid

- Masinõpe
  - Anname arvutile palju andmeid ja näite, mida tahame saada, õpib ise
  - Vaja läheb: suuri korpusi, inimmärgendatud korpuseosa, parameetrite muutjat, hindamist
- Juhendamata õpe
  - Tegelikult õpib arvuti ka ilma näiteta päris hästi
  - Vaja läheb: peamiselt korpusi, inimpüstitatud eesmärki, hindamist

Töövõimalused

Näide aastast 1991: artiklite sissejuhatuste põhjal -> magneesiumi aitab migreeni vastu

Stress associated with Migrain headaches

Stress leads to loss of magnesium

Calcium channel blockers prevent some migrain headaches

Magnesium is a natural calcium channel blocker

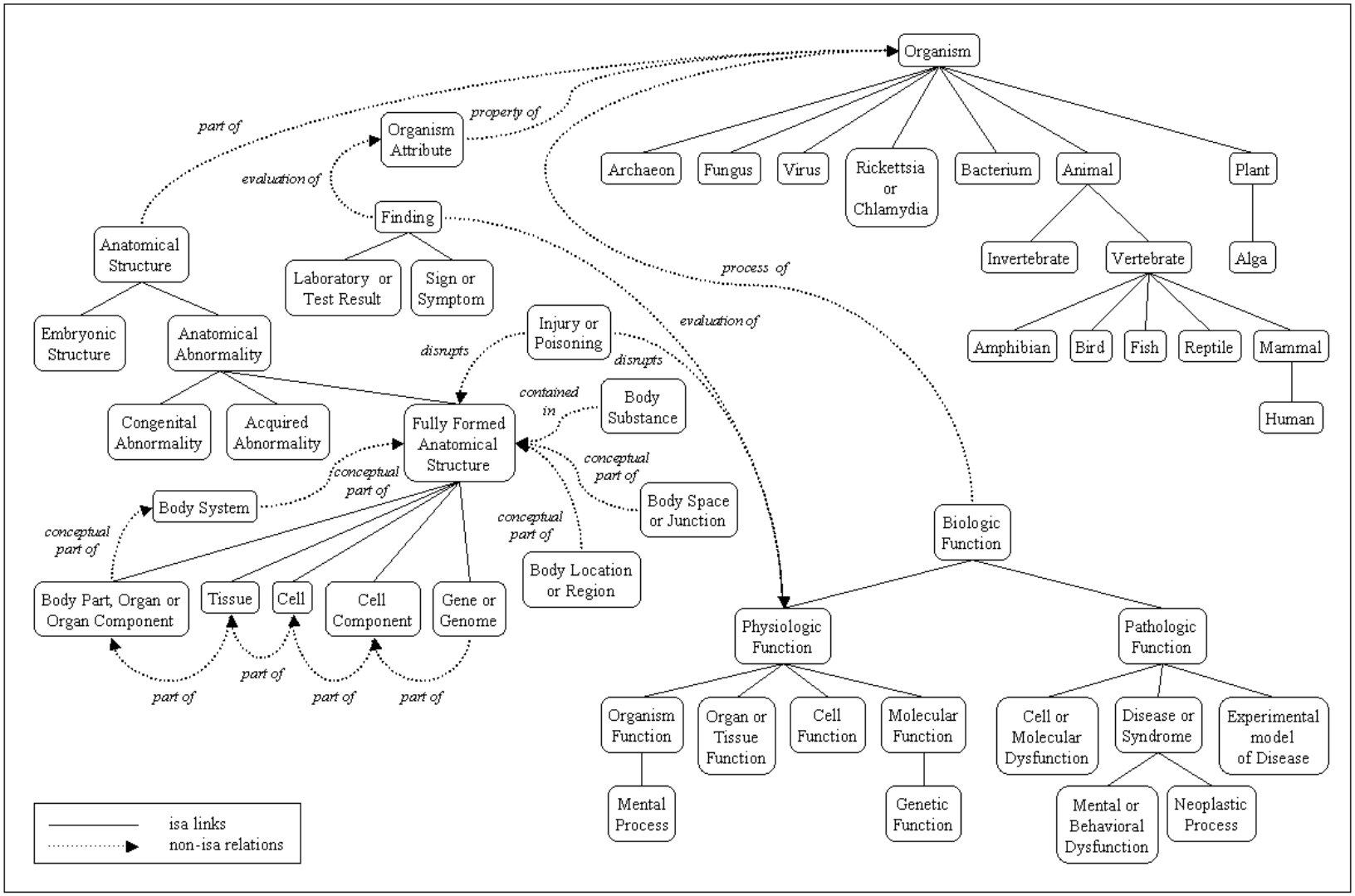
SCD implicated in migrain headaches

High levels of magnesium inhibit SCD

Migrain patients have high platelet aggregability

Magnesium can suppress platelet aggregability

Tekstikaeve: võrgustik



Pilt pärineb: aine “Text Mining”, Fabio Rinaldi, 2007

Keeletehnoloogia välismaal

- Microsoft, Google jm – otsingu optimeerimine, tõlge
- Daimler – autode hääljuhtimine
  - “Nihuta seda peeglit natuke ülespoole”
- Novartis – ravimifirma
  - Uuri bioloogide, arstide ja geneetikute artikleid ning anna vihjeid, mida tasuks uurida

Keeletehnoloog Eestis

- Ettevõtlus ei tasu ära
  - Projektid, mil eeltöö tehtud -> Filosoft
  - Lokaliseerimine jm lisaks -> Tilde
  - Eurorahad -> Tilde
- Tartu Ülikool
  - Spetsialiseerumine keeletehnoloogiale
  - Spetsialiseerumine arvutuslingvistikale
- Tallinna Tehnikaülikool
  - Foneetika ja kõnetehnoloogia laboratoorium
- Eesti Keele Instituut

Projektid

- Projekt mingi finantsallika alla
  - Riiklik programm
  - Sihtfinantseerimine
  - Euroopa liidu programm
- Iga-aastane aruandlus ja uuesti raha taotlemine
- Inimesed võetakse tööle täpselt projekti kestvuseks
- Teadustöö eesmärgid seatakse täpselt projekti kestvuseks

Keeletehnoloogi igapäevatöö

- “Tegelik töö” arvuti taga
- Publitseerimine
- Konverentsid
- Õppetöö
  - Mitte tingimata kohustuslik, aga…
- Administreerimine

Grammatikakorrektor

Morfoloogiline analüüs
Morfoloogiline ühestamine

Grammatikakorrektori reeglid

Kitsenduste grammatika: teooria

Kitsenduste grammatika: parser

CG reegel

ellepärast et, selleks et, selle asemel et, nii et; siis kui, enne kui, juhul kui, nii palju kui, samal ajal kui jms puhul on koma asend kõikuv. Kui rõhutatakse põhjust, otstarvet vms, siis paigutatakse sõnad *sellepärast, selleks* vms pealausesse ning koma sidesõna *et, kui* vms ette. Muidu on koma ühendsidendi ees. Pealausele eelneva kõrvallause algul on ükskõik, kas ühendi osade vahele koma panna või mitte. Tarvis seda koma pole.

ii et püsiühendites võib koma üldse panemata jätta.

SELECT (@OK) (0 ("et+0")) (-1 ("nii+0")) (NOT -2 Eraldaja);# nii et
SELECT (@ERR) (0 ("et+0")) (-1 (Com)) (-2 ("nii+0")) (-3 (Com)); #,nii, et
SELECT (@OK) (0 ("et+0")) (-1 ("nii+0")) (-2 Eraldaja); #, nii et
SELECT (@OK) (0 ("et+0")) (-2 ("nii+0")) (-1 (Com)); # nii, et

+ CG reegel

Müüsin vana auto maha hetkel mil oleks pidanud riskigrupp langema 0.77-0.60-le aga uut autot kindlustama minnes tõusis hoopis 0.87.

Mina ei käinud eile vaatama oma korter.

Muul juhul peab koma nõudva sidesõna ees peab olema koma.

Spelleri integreerimisega veebibrauserisse muutub ka meilide õigekiri praemaks.

Artiklid

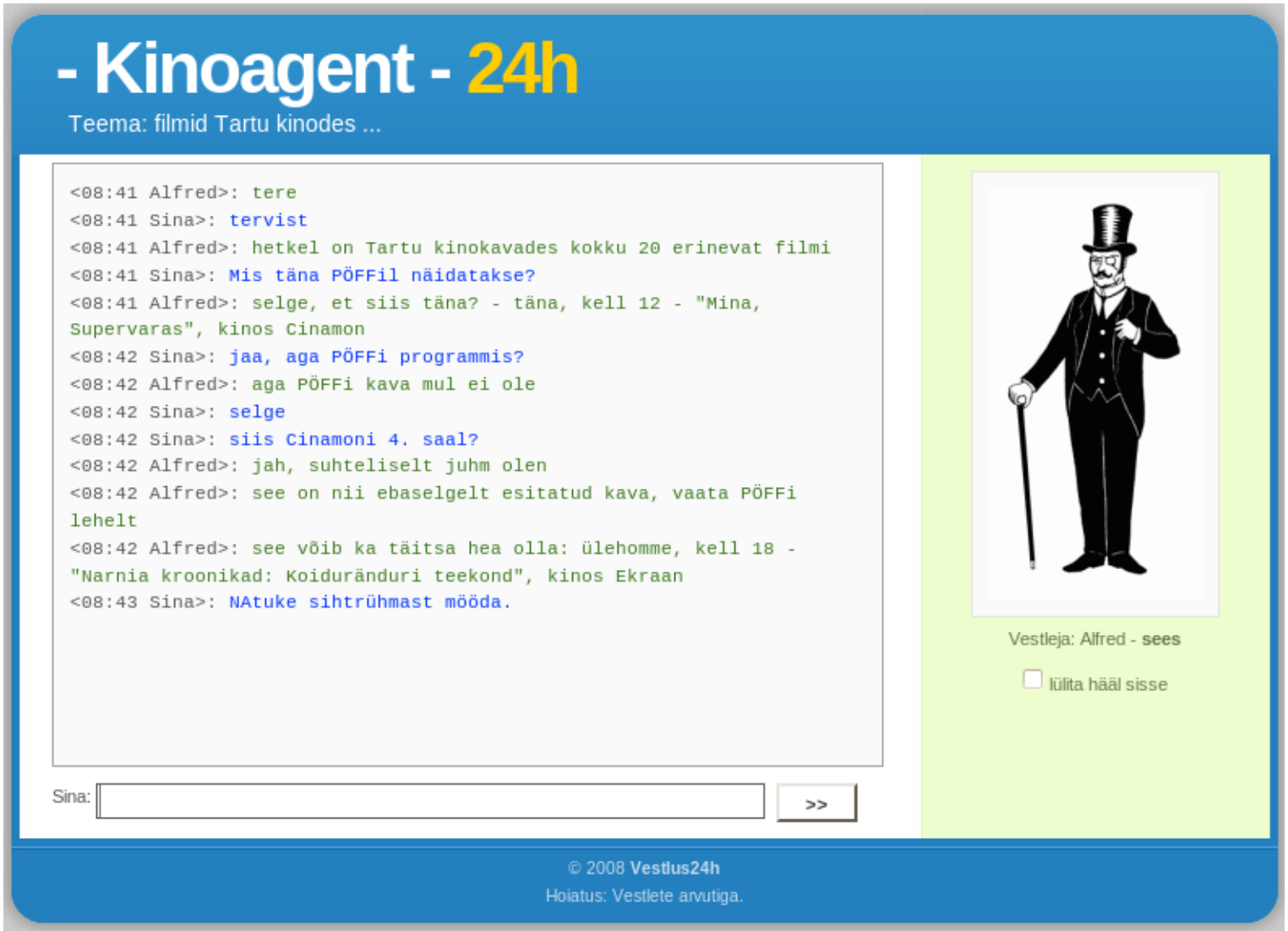
- Publikatsioonide taseme ja arvu põhjal hinnatakse
  - Ülikooli teadust
  - Projektide tulemusi
  - Keskuste taset
  - ...
- Keeletehnoloogia – pearõhk konverentsidel
  - Konverentsikogumikud

Konverentsid

- Ettekanne
  - ~15 min tutvustust mitme kuu/aasta tööle
  - Ettevalmistus bakatöö kaitsmiseks, levinud moodus midagi tutvustada – piisav ettevalmistus
- Poster
  - ~1,5 tundi postri ees seismist ja seletamist
  - Ettevalmistus valitud ainetes
- Töö vs puhkus
  - Paralleelsessioonid
  - Tavapäraselt 9-18

Dialoog eesti keeles

http://www.dialoogid.ee/kinoagent/

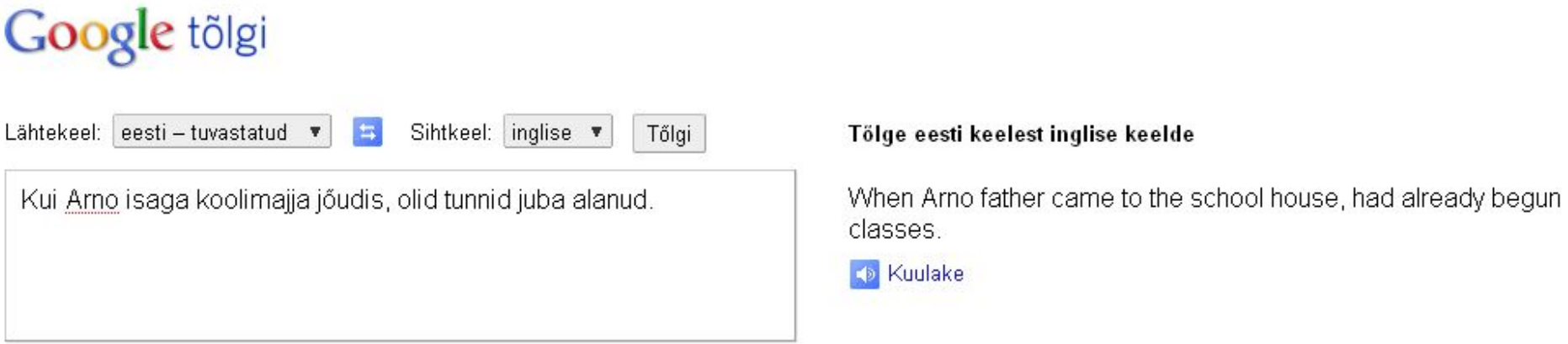


Masintõlge

http://masintolge.ut.ee/

Masintõlge

http://translate.google.com/



Lisamaterjale

Sõnaraamatud ja korpused:

http://www.keeleveeb.ee/

Keeletehnoloogia uurimisrühma lehed

http://www.ut.ee/~koit/KT/

http://www.cl.ut.ee/

Küsimusi?