

TARTU ÜLIKOOL

Eestikeelne infodialoog arvutiga (2006-2008)

Intelligentne kasutajaliides andmebaasidele (2009-2010)

Mare Koit

TARTU ÜLIKOOL

Põhitäitjad

- Mark Fiel,
- Olga Gerassimenko,
- Riina Kasterpalu,
- Krista Mihkels,
- Siiri Pärkson,
- Andriela Rääbis,
- Margus Treumuth.

TARTU ÜLIKOOL

Finantseerimine

- 2006 350 000
- 2007 350 000
- 2008 550 000
-
- 1 250 000
- 2009 540 000
- 2010 470 000
-
- 1 010 000

TARTU ÜLIKOOL

Eesmärk

- Luua kasutajaliides, mis
 - võimaldab häälestamist erinevatele ainevaldkondadele ja seostamist erinevate andmebaasidega,
 - kasutaja saab andmebaasiga suhelda eesti keeles ja inimestevahelise suhtluse reeglite kohaselt
 - kasutatav ka võlur Ozi režiimis (et koguda andmeid häälestamiseks uuele ainevaldkonnale)
- Selleks
 - (märgendatud) dialoogikorpus
 - dialoogiaktide märgenduskeem
 - dialoogiaktide märgendamise tarkvara
 - dialoogiaktide automaatse tuvastamise meetodid
 - eestikeelse infodialoogi juhtimise mudel

TARTU ÜLIKOOL

Tööhüpotees

J. Allen: praktilise dialoogi hüpotees (2001)

- Erinevates ainevaldkondades leiduvad ühised minimaalsed baastunnused, mis
 - tagavad põhilise suhtlusfunktsionaalsuse,
 - samal ajal säilib ainevaldkonna-spetsiifilise info kättesaadavus ja antava info kasulikkus.

TARTU ÜLIKOOL

Jäik vs paindlik dialoog arvutiga

Margus Treumuth

- Vooruvahetus jäik
 - Reisiagent
 - Teatriagent
- Vooruvahetus paindlik
 - Kinoagent: info Tartu kinodes linastuvate filmide kohta <http://www.dialoogid.ee/kinoagent>
 - Zelda: hambaraviinfo

TARTU ÜLIKOOL

Kuidas liides töötab?

Tervitus
KORRATA:

- Üldine info ainevaldkonna kohta, oodates kasutaja sisendit
- Kasutaja sisendi
 - morfoloogiline analüüs (ESTMORF)
 - õigekirjakontrolli, vigade parandamine
 - lihtne semantiline analüüs
 - ainevaldkonna võtmesõnade ja fraaside tuvastamine
 - ajaväljendite tuvastamine
 - pärisnimede tuvastamine
 - vajadusel sõnajärjestuse muutmine (s.t genereeritakse kõik permutatsioonid, et sobitada kasutaja sisend semantiliselt tunnusega).
- SQL-päring andmebaasile
- Vastus eesti keeles, kasutades lausemalle ja vajadusel morfoloogilist sünteesi
 - Valikuliselt tekst-kõnesüntees

TARTU ÜLIKOOL

Võlur Oz



TARTU ÜLIKOOL

Dialogikorpus

<http://epo.it.da.ut.ee/~koit/Dialogi/EDIC.html>

Olga Gerassimenko, Riina Kasterpalu, Krista Mihkels, Andriela Rääbis; Siiri Pärkson, Margus Treumuth

- Dialogiaktidega märgendatud** dialooge: 1220 (245 000 tekstisõna)
 - Suulised inimestevahelised 1146
 - Võlur Ozi meetodil kogutud 22 (2001.a) + 52 (2009.a)
- Lisaks
 - Inimese ja arvuti vahelised (logifailid)
- TÜ dialoogiaktide tüpoloogia (Tiit Hennoste, Andriela Rääbis)
 - 126 dialoogiakti, jagunevad 19 klassi (naaberpaariaktid ja üksikaktid, dialoogi juhtimise aktid ja infoaktid)

TARTU ÜLIKOOL

Dialogikorpus

- Dialoogiaktide märgendusprogramm
 - Evely Vutt + Maret Valdisoo programm (keskkond) dialoogiaktide käsitsi märgendamiseks
 - Mark Fīel, Taavet Kikas dialoogiaktide automaatne tuvastamine eestikeelsetes dialoogides
 - Tehisnärivõrgud (mitmekihiline tajur, rekurrentne kihiline tehisnärivõrk)
 - Otsustuspuud
 - Tõenäosuslikud sufiksipuud
 - Naaiivne Bayes: täpsus 62%

TARTU ÜLIKOOL

Dialogikorpus

- Uuritud ja arvesse võetud intelligentsetes liideses
 - Infotelefonikõnede struktuuri
 - Interneti-suhtluse struktuuri
 - Rituaalsete aktide (tervitused, hüvastijätud, tänamised jm) ja infoaktide (soovid, küsimused, info andmised jm) väljendamist eesti keeles
 - Parandussekventse (paranduse algatus, läbiviimine ja vastuvõtmine)

TARTU ÜLIKOOL

Dialogikorpuse tööpink

Margus Treumuth <http://www.dialoogid.ee/dialoog>

Võimaldab:

- valida alamkorpust ja dialoogi,
- näidata dialoogi kulgu ajateljel,
- koostada sõnavormide sagedustabelit,
- teha morfoloogist analüüsi (ESTMORF),
- otsida dialoogiakte nime ja osalejatunnuse järgi,
- dialoogiakte poolautomaatselt märgendada (Mark Fīel).

Tulemused: kokkuvõte

- Veebipõhine tarkvara eestikeelse dialoogi pidamiseks (intelligentne liides andmebaasidele).
- Liidest saab häälestada uutele ainevaldkondadele ja siduda erinevate andmebaasidega.
 - Kasutaja pöördub andmebaasi poole eesti keeles ja saab vastuseks adekvaatset, tõest infot.
- Liideses on lõimitud eesti keele automaattöötuse vahendid: morfoloogiline analüüs ja süntees, õigekirjakontroll ja vigaste vormide korrigeerimine, nimega üksuste (ajaväljendid ja pärisnimed) tuvastamine, tekst-kõnesüntees.
- Liidest on testitud kinoinfo ja hambaraviinfo andmebaasidega.

Tulemused: kokkuvõte

- TÜ dialoogikorpuses on 1220 *dialoogiaktidega märgendatud* dialoogi (245 000 tekstisõna).
- Korpuse on ligipääsetav veebis (parooliga kaitsitud).
- Märgendamisel on kasutatud TÜ dialoogiaktide tüpoloogiat (Hennoste, Rääbis).
- Korpuse analüüsimiseks on välja töötatud tarkvara dialoogikorpuse tööpink.