Vahendid teksti mitmekihiliseks märgendamiseks
(rakendatuna Koondkorpusele)

Kadri Muischnek

Taust

Projekt ühendab endas lõppenud EKKT projektide

„Eesti keele koondkorpus“,

„Eesti keele sõltuvusgrammatika arendamine ja osaliselt mittekorrektse eestikeelse teksti morfoloogiline ühestamine ja süntaktiline analüüs“

ning “Korpusepäring Keeleveebis“

tulemuste edasiarendamist

Eesmärgid

Põhieesmärk: koondada senised korpuse märgendamiseks kasutatud tarkvaraprototüübid ühtseks standardiseeritud programmide koguks ja nende abil muuta eesti keele Koondkorpus (240 miljoni sõna suurune tänapäeva kirjaliku keelekasutuse korpus) mitmetasandiliselt (morfoloogiliselt, süntaktiliselt, semantiliselt) märgendatud korpuseks

Kaks **alameesmärk**i:

Esiteks: luua stabiilselt töötavad versioonid olemasolevatest tarkvaraprototüüpidest

Teiseks: Koondkorpuse täiustamine ja selle kasutusvõimaluste avardamine

Täitmine: morfoloogia ja süntaks

Algseis:

Morfoloogia: stabiilselt töötav analüsaator ja ühestajad olid olemas; nende väljundit saab kasutada teiste märgenduskihtide loomiseks.

Vajalik morfoloogilise ühestamise protsessi täiustamine; erinevate analüsaatorite ja ühestajate kombineerimise võimaldamine

Süntaks: oli olemas kitsenduste grammatikal põhinev pindmine süntaksianalüsaator. Pindmise analüüsi põhjal leiab süvasüntaksianalüsaator ka lause sõltuvusstruktuuri (lausepuu või osalised puufragmendid), kuid see analüsaatori versioon ei olnud stabiilne.

Vajalik töökindla sõltuvussüntaktilise analüsaatori loomine ja suuremahuline korpuse süntaktiline märgendamine

Täitmine: morfoloogia ja süntaks 2

Tehtud:

Tarkvara täiustamise osas:

integreeritud kitsenduste grammatikal põhinevad morfoloogiline ühestaja ja CG süntaksianalüsaator

süntaksianalüsaator on kohandatud ka statistilise morfoloogilise ühestaja (*t3mesta*) väljundile, st süntaksianalüsaator töötab nüüd mõlema eesti keele jaoks olemasoleva morfoloogilise ühestaja väljundiga.

CG süntaksianalüsaator (koos integreeritud morf ühestajaga) on adapteeritud EKI morf analüsaatori väljundile, st töötab nüüd mõlema eesti keele jaoks olemasoleva morfoloogiaanalüsaatori väljundiga

täiustatud CG sõltuvussüntaktilise analüsaatori reegleid; käsitsi märgendatud sõltuvuspuude pank, ca 80 000 sõna

Täitmine: morfoloogia ja süntaks 3

On läbi viidud järgmised katsed:

1) reeglipõhise ning statistilise morfoloogilise ühestaja väljundi süntaksianalüsaatori sisendiks sobivuse võrdlus
2) kahe ühestaja, st CG reeglipõhise ja statistilise *t3mesta* väljundile rakendatud pindsüntaktilise analüüsi tulemuste võrdlus
3) Süntaksianalüüsi (koos integreeritud morf ühestamisega) tulemused EKI vs Filosofti morfanalüsaatorilt saadud tulemuste puhul (sellest tuleb ettekanne HLT-l)

Täitmine: morfoloogia ja süntaks 4

Korpuse märgendamine:

Koondkorpusele lisatud osalausepiirid; selle kohta vt täpsemalt artiklit ERÜ 2011. aastaraamatus NÄIDE

Koondkorpuses katseliselt märgendatud minimaalsed NP-d, nt

ilus müts, tütre müts

2012. aasta lõpuks märgendatakse morfoloogiliselt kogu Koondkorpus, kasutades selleks reeglipõhist (CG – kitsenduste grammatika) ühestajat. Pindsüntaktiliselt (CG) märgendatakse Tasakaalus korpus (Koondkorpuse 15 miljoni sõnaline allosa).

Käsitsi märgendatud sõltuvuspuude pank, praegu ca 80 000 sõna

Täitmine: praktiline semantika

Praktiline semantiline analüüs - teatud semantiliste klasside, nt. nimede, ajaväljendite, teatud liiki asjade (nt autode, elukutsete, keemiliste ainete) esinemisjuhtude märgendamine tekstis.

Tehtud: loodud programm nime- ja numbriüksuste ning ajaväljendite märgendamiseks ja sellega märgendatud Koondkorpus

Semantiline märgendus on lisatud ka Keeleveebi kaudu kasutatavale Koondkorpuse versioonile.

NÄITED

Täitmine: praktiline semantika 2

Kuidas? – reeglipõhised programmid, arvestavad sõnade käändevormide sagedusjaotust antud tekstis, sõna kuulumist teatud loendisse ja sõnade lokaalset konteksti jms

Täpsus ja saak muude nime- ja numbriüksuste eraldamisel ja liigitamisel u 80%;

ajaväljendite eraldamisel ja liigitamisel f-mõõt 89%

Tekstiliigi automaatse tuvastamise eeltööd

Milleks? Tekst võib oma tüübilt olla formaalne või mitteformaalne; kirjakeelne või mitte. Teksti tüübi teadmine on oluline selleks, et programmid saaksid seda arvestada.

nt kui tekstis suurtähti ei kasutata, siis tuleb lausete ja nimede eristamiseks kasutada muid tunnuseid;

nt teadustekstis tuleb leksikonist puuduvate sõnade analüüsiks kasutada teistsuguseid heuristikuid kui mitteformaalse netikeele puhul.

Tekstiliigi automaatse tuvastamise eeltööd 2

Tehtud: Tasakaalus korpuse põhjal on koostatud sõnavormide ja lemmade sagedusloendid allkorpuste kaupa, vt lähemalt http://www.cl.ut.ee/ressursid/sagedused1

Valmimas:

grammatiliste kategooriate sagedusloendid

Kollokatsioonide tuvastaja

www.rabauti.ee/clc

Nüüd saab Tasakaalus korpusest otsida **osalauses** esinevate sõnavormide või lemmade koosesinemisi.

Kollokatsioone saab otsida kolmel viisil:

1) teatud lemma olulisi kollokaate sõnavormidena

2) teatud lemma olulisi kollokaate lemmadena

3) teatud sõnavormi olulisi kollokaate sõnavormidena

Nii sisestava lemma või sõnavormi kui ka otsitavate kollokaatide ringi saab piirata nende sõnaliigilise kuuluvusega

NÄIDE

Koondkorpuse parandamine ja edasiarendamine

Süstemaatiliste lausestusvigade parandamine -> parem morf -> paremad järgmised analüüsietapid

Täiendatud Koondkorpuse internetikeele e uue meedia keelekasutuse allosa: kogutud 4 miljonit sõna jututubade tekste http://www.cl.ut.ee/korpused/segakorpus/uusmeedia/jututoad2

Publikatsioonid

Tiina Puolakainen (2012). How Does the Choice of Morphological Analyser Influence the Quality of Syntactical Analysis. Proceedings of Baltic HLT 2012

Kaalep, Heiki-Jaan; Muischnek, Kadri (2012). Osalausete tuvastamine eestikeelses tekstis kui iseseisev ülesanne. Eesti Rakenduslingvistika Ühingu aastaraamat (55 - 68).

Kaalep, Heiki-Jaan; Muischnek, Kadri (2012). Robust clause boundary identification for corpus annotation. Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12) (1632 - 1636).

Täitjad ja finantseerimine

Vastutav täitja: Kadri Muischnek

Teised põhitäitjad: Kaili Müürisep (praegu lapsehoolduspuhkusel), Tiina Puolakainen, Riin Kirt, Raigo Kodasmaa, Dage Särg, Katrin Tsepelina, Kristel Uiboaed, OÜ Filosoft

Finantseerimine 2011: 97500

Finantseerimine 2012:85000