TARTU ÜLIKOOL
MATEMAATIKA-INFORMAATIKATEADUSKOND
Arvutiteaduse instituut
Informaatika eriala

Kristi Zirk

# Reeglipõhine ühestaja eesti keele jaoks

Bakalaureusetöö (6 EAP)

Juhendaja: teadur Neeme Kahusk

Autor: …… "….." ……... 2013

Juhendaja: ……. "….." ……... 2013

Lubada kaitsmisele

Professor: …… "….." ……... 2013

TARTU 2013

# Sisukord

**Sissejuhatus** ..... **3**

**1. Sõnatähenduse ühestamine** ..... **4**

1.1. Sõnatähenduse ühestamisest üldiselt ..... 4

1.2. Sõnatähenduste ühestamise alusleksikon ..... 5

1.2.1. TEKsaurus ..... 5

1.3. Eesti sõnatähenduste ühestatud korpus ..... 6

1.3.1. Sõnatähenduse ühestamise analüüsi etapid ..... 8

1.3.2. Käsitsi sõnatähenduste ühestamine ..... 10

1.4. Automaatne sõnatähenduste ühestamine ..... 11

1.5. Ühestamise meetodid eesti keele jaoks ..... 12

1.5.1. Reeglipõhine ühestamine ..... 12

1.5.2. Statistiline ühestamine ..... 12

1.5.3. Hetkeolukord ..... 13

**2. Programmi ülevaade** ..... **15**

2.1. Reeglid ja nende formaliseerimine ..... 15

2.2. Kasutatavad tehnoloogiad ja algoritmid ..... 16

2.3. Kasutusjuhend ..... 16

2.3.1. Programmi käivitamine ..... 16

2.3.2. Programmi sisend ja väljund ..... 16

2.4. Programmi testimine ja töökiirus ..... 16

**Kokkuvõte** ..... **18**

**Summary** ..... **19**

**Kirjandus** ..... **20**

**Lisad** ..... **21**

**Lisa 1. CD lähtekoodiga** ..... **21**

# Sissejuhatus

Sõnatähenduste ühestamine (STÜ, ingl *word sense disambiguation*) on semantilise ühestamise üks allülesandeid, seega on ta üks olulistest ülesannetest, mis tuleks lahendada selleks, et keeletehnoloogilised rakendused tuleksid toime loomuliku keelega. STÜ puhul omistatakse sõnale just see tähendus, mis tuleneb tema kontekstist. (Agirre, Edmonds 2006) Näiteks on vaja masintõlkesüsteemis eristada nimisõna *naine* tähendust – kas tegemist on abielunaisega (ingl *wife*) või sellisest soost inimesega, kes on võimeline lapsi saama (ingl *woman*). Sama probleem tekib sõna *tee* eristamisel – kas tegemist on teega kui joogiga (ingl *tea)*, teega nagu rajaga (ingl *path*) vms. Selle ülesande lahendamiseks on lisaks tekstile vaja ka leksikoni, kus oleksid määratud sõnade tähendused. Sobivaks leksikoniks STÜ jaoks on WordNet. Eesti keele korral kasutatakse Eesti Wordneti ehk TEKsaurust[1].

Töö koosneb kahest peatükist. Käesoleva töö esimene, teoreetiline osa annab ülevaate sõnatähenduste ühestamisega seotud mõistetest ja protsessidest. Kuna Eestis on võetud peamiseks STÜ leksikoniks TEKsaurus, siis on antud ka sellest lühike ülevaade. TEKsauruse põhjal toimub eesti keele puhul ka nii automaatne kui ka käsitsi sõnatähenduse ühestamine. Automaatne sõnatähenduse ühestaja eesti keeles on **semyhe**, mis on kirjutatud Kaarel Kaljuranna poolt. Peamiselt on sõnatähendusi märgendatud käsitsi. Peatüki lõpu poole tuleb juttu erinevatest ühestamisel kasutatud meetoditest ja antakse ülevaade hetkeolukorrast.

Praktilise osa eesmärgiks on formaliseerida olemasolevad sõnatähenduste ühestamise reeglid ja luua programm, mis kasutaks neid reegleid sõnatähenduste märgendamiseks korpuses. Sõnatähenduste ühestamise reeglid on formaliseeritud praegu eestikeelsete lausetena, mis on abiks leksikograafidele õige sõnatähenduse määramisel. Hetkel on reegleid umbes 100 ja need kirjeldavad sagedamini esinevaid polüseemseid nimisõnu. Programmile antavas sisendfailis märgendatakse tekstis olevad sõnad vastavalt semantikale. Märgendamisel lisatakse väljundfaili sõna lemma (sõnaraamatu vorm) ja tähendusnumber. Olemasolevad reeglid on kirjutatud Kadri Vare poolt ning need on loodud selleks, et aidata leksikograafe, kes märgivad tähendust.

Lisana on esitatud programmi lähtekood (CD plaadil).

[1] http://www.cl.ut.ee/ressursid/teksaurus/

# 1. Sõnatähenduse ühestamine

Sõnatähenduse ühestamine on keeletehnoloogia ülesanne, mille eesmärgiks on otsustada, millises tähenduses sõna kasutatakse etteantud kontekstis. STÜ algab eesti keeles juba morfoloogilise analüüsi tasandil ja on semantilise analüüsi üks esimestest etappidest. (Kerner 2007, 6) Kuigi morfoloogiline analüüs ja ühestamine on sõnatähenduse ühestamise eelduseks, ei käsitleta (käesolevas töös) seda rohkem, kui üldiste ühestamispõhimõtete seletamiseks ja näitlikustamiseks.

Sõnatähenduste ühestamist peetakse loomuliku keele töötlemise juures üheks peamiseks probleemiks (Ide,Veronis 1998). Seda vajavad näiteks masintõlge, infootsing, kõnetöötlus jne.

## *1.1. Sõnatähenduse ühestamisest üldiselt*

Erinevates kontekstides võib ühel sõnal olla erinevad semantilised interpretatsioonid. Selleks on kolm võimalust, milleks on homonüümia, polüseemia ja ebamäärane tähendus. Kui sõnal on üks tähendus, mis on piisavalt üldine, siis on sõnal ebamäärane tähendus. Kui sõnal on rohkem kui üks võimalik tähendus, on tegu mitmetähendusliku sõnaga. Homonüümiaga on tegu siis, kui kaks vormi-tähenduse paari toovad kaasa kaks erinevat lekseemi, millel on juhuslikult sama hääldus/kirjakuju ja vastavaid lekseeme nimetatakse homonüümideks. Sellised on näiteks sõnad *tint, palk* ja *aru* – samakujulised, kuid eri tähendusega sõnad. Homonüümid jagunevad omakorda homofoonideks ja homograafideks. Homofoonid on sõnad, mida hääldatakse ühtemoodi, kuid mille kirjapilt võib erineda, sellised on näiteks sõnad *baar* ja *paar*. Homograafid on sõnad, mida kirjutatakse ühtemoodi, aga mille hääldus võib erineda, sellised sõnad on *tulp (tulba)* ja *tulp (tubli)*. (Õim, 2012) Polüseemia on aga nähtus, mille puhul ühel sõnal on mitu üksteisega tihedalt seotud tähendust. Üheks selliseks sõnaks on *klaas*, millele TEKsaurus annab neli erinevat tähendust. Vastavalt TEKsaurusele võib klaas olla tahke aine, asi, jooginõu või mahumõõt.

Väidetavalt on homonüümia ajalooline või juhuslik, polüseemia aga võib tuleneda keele ökonoomsusest, kasutades ära juba keeles olemas olevaid sõnu (Kerner 2007, 6). Mitmetähenduslikkus on eesti keeles levinud, näiteks on nimisõnal *asi* (vastavalt TEKsaurusele) 11 erinevat tähendust, verbil *käima* 24 tähendust. Erinevate keeletehnoloogiliste rakenduste jaoks on vaja ühestada nii polüseemseid kui homonüümseid sõnu.

Eesti keele puhul on STÜ jaoks vaja kõigepealt teha morfoloogiline analüüs ja ühestamine. Kuigi lauses „Ta hakkas kolima“ puhul on võimalik ainult verbi analüüs

’*koli+ma // _V_ ma //’*, siis ’koli’ puhul on võimalik nii verbi kui nimisõna analüüs, vastavalt *’koli+0 // _V_ main imper //’* ja *’koli+0 // _S_ sg g, sg n, sg p//’*.

Sõnatähenduse ühestamise vastu hakati huvi tundma juba alates 1950ndatest aastatest. Sõnatähenduse ühestamine on vahendava ülesandega, seda vajavad näiteks:

- masintõlge – tõlkida tuleks õige tähendus. Kas sõna *naine* tuleks tõlkida inglise keelde kui *woman* või kui *wife* (abielunaine)? (Kerner 2007, 7)
- info-otsing, mille juures on vajalik eemaldada sellised tähendused, mis ei ole vastaval otsingul vajalikud;
- grammatiline analüüs – lauseliikmete määramiseks juhul, kui sõna tähenduse teadmina teatud kontekstis on oluline; (Kerner 2004, 5)
- kõnetöötlus – samamoodi kõlavate sõnade eristamine. Näiteks *tulp:tulba* ja *tulp:tulbi* (Kerner 2007, 7).

## *1.2. Sõnatähenduste ühestamise alusleksikon*

Sõnatähenduste ühestamine käib mingi etaloni alusel. Etalonis on eristatud sõnade tähendused teatud tunnuste võrdlemise ja välistamise teel. Sel moel piiritletud sõnatähenduste kogum on sõnatähenduste ühestamise alusleksikoniks, millest tekstisõna kontekstiga parimini sobivat tähendust otsitakse. (Kerner 2007, 7) Antud töös on alusleksikonina kasutatud eesti keele *wordnet*-tüüpi tesaurust – TEKsaurus.

### 1.2.1. TEKsaurus

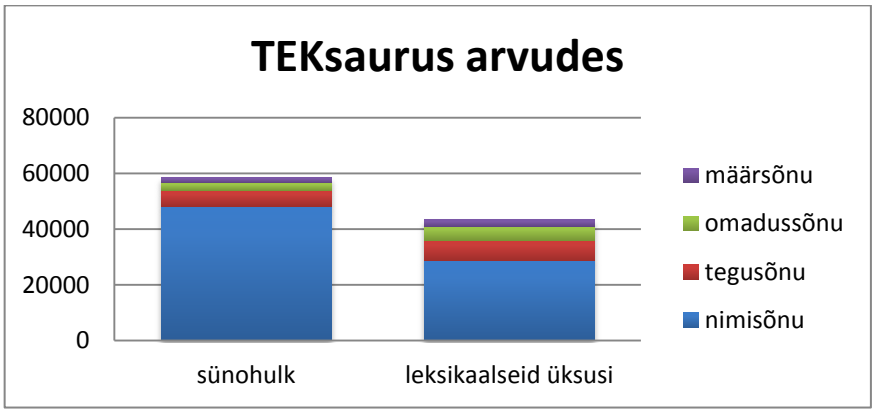
Tartu Ülikooli eesti keele tesaurus ehk TEKsaurus (Eesti Wordnet) loodi 1998. aastal EuroWordNeti[2] projekti käigus, kus sarnaste põhimõtete järgi koostati kaheksa erineva keele tesaurused. TEKsaurus on keele leksikaal-semantiline andmebaas, kus on lisaks sõnade tähenduste eristamisele fikseeritud ka tähendustevahelised seosed. Eesti Wordneti muudab väärtuslikuks selle mitmekeelsus, sest ta on seotud EuroWordNetiga, mille kaudu on Eesti Wordnetis olevad mõisted ühendatud ingliskeelse keeltevälise indeksi (InterLingual Indexi (ILI-link)) abil ka teiste keelte tesauruste mõistega. (Orav 2011, 11) WordNetis sisalduvad ainult täistähenduslikud sõnad: nimisõnad, tegusõnad, omadussõnad ja määrsõnad. Eesti Wordnet’i peamised tegijad on olnud Neeme Kahusk, Heili Orav, Kadri Vider jt, järgides Princetoni WordNet’i ja EuroWordNet’i põhimõtteid (Lõo 2010, 10).

[2] http://www.illc.uva.nl/EuroWordNet/ (12.03.2013)

Sünohulk ehk sünonüümirida (ingl. *synonym set, synset*), on WordNet'i elementaarosake, mille moodustavad ühte mõistet (*concept, meaning*) väljendavad sünonüümsed (sama tähendusega) sõnad ja sõnaühendid. Termin sünohulk on loodud sellepärast, et erinevalt sünonüümisõnastiku sünonüümireast võib sünohulk olla ka üheliikmeline.[3] Sünohulgad on omavahel seotud semantiliste suhete kaudu. Semantilisi suhteid on erinevaid liike ja mõnikord nimetatakse mõningaid neist ka leksikaalseteks suheteks.

Eesti Wordnet'is on sünohulki üle 58 000, millest on umbes 48 000 nimisõnad, 5700 tegusõnad, 2800 omadussõnad ja 1800 määrsõnad. Semantilisi suhteid on üle 174 000 ja ILI-suhteid umbes 63 000.

Kõigist tesaurusest esitatud leksikaalsetest üksustest (sõnades ja sõnaühenditest) on 80% ühe tähendusega, seega mitmetähenduslikke üksusi on 20%. Nendest mitmetähenduslikest (kahe ja enama tähendusega) üksustest on 74% omakorda kahe tähendusega sõnad, 15% kolme tähendusega ja 11% sõnu omab rohkem kui kolme tähendust, sealhulgas 0,08% sõnu omab rohkem kui 10 eri tähendust.




Tabel 1 - TEKsauruse andmed[4]

## *1.3. Eesti sõnatähenduste ühestatud korpus*

Eesti STÜ korpus[5] koosneb kahest osast: baaskorpusest ja liikumislausete korpusest, neist mõlemates on tekstid käsitsi märgendatud. Vaatame lähemalt baaskorpuse materjali ja andmeid.

[3] http://www.cl.ut.ee/ressursid/teksaurus/ (18.03.2013)
[4] http://www.cl.ut.ee/ressursid/teksaurus/index.php?lang=en (02.05.2013)
[5] vaata ka http://www.cl.ut.ee/korpused/semkorpus/ (21.03.2013)

Ühe sõna analüüsi tulemus on sarnane sellele, mida on kirjeldatud morfoloogiliselt ühestatud korpuse[6] juures, samuti on info semantika ehk sõnatähenduse kohta: (Kerner 2007, 10)

*sõna  tüvi + lõpp // morfoloogiline analüüs // semantiline info*

- <sõna> on sõna sellisena, nagu ta algselt esines
- <tüvi> on lemma e. algvormi tüvi: käändsõnadel ainsuse nimetav (kui seda olemas ei ole, siis mitmuse nimetav), pöördsõnadel ma-infinitiivi tüvi ilma (ma-lõputa)
- <lõpp> on sõna lõpp, kusjuures mitmuse tunnus on temaga liitunud (nagu seda on käsitletud ka Ülle Viksi "Väikeses vormisõnastikus"); partikkel GI/KI, kui ta esineb, on lihtsalt lõppu "kleepunud"; ka juhul, kui sõnal ei saagi lõppu olla (nt. hüüdsõnal), pannakse sõnale lõpp - nn. null-lõpp
- <morfoloogiline analüüs> on üks variantidest, mis on kõik esitatud morfoloogiliste kategooriate tabelis.
- <semantiline info> on TEKsauruse sünohulk koos vastava tähendusnumbriga.[7]

Kui on tegemist liitsõna või tuletisega, siis:

- Tüvi on eristatud eelnevast komponendist '_' märgiga;
- Lõpp on eristatud eelnevast komponendist '+' märgiga; nn. null-lõpp ongi '+0'
- Sufiks on eristatud eelnevast komponendist '=' märgiga. Sufiksite märkimine ei ole järjekindel: märgitakse ainult teatud hulka produktiivseid sufikseid.
- Lemmatüvi leitakse ainult viimase parempoolse komponendi alusel

Mitmesõnalised nimed on sellisel kujul:

New Yorgis New York+s //_S_ prop sg in //

Omaette ridadel asuvad märgendid <s> ja </s> tähistavad lause või pealkirja algust ja lõppu; mõnedes failides esinevad  ja  tähistavad lõigu algust ja lõppu. [8]

[6] www.cl.ut.ee/korpused/morfkorpus (21.03.2013)
[7] www.cl.ut.ee/korpused/morfkorpus (10.04.2013)
[8] www.cl.ut.ee/korpused/morfkorpus (10.04.2013)

Eesti keele baaskorpuses on kokku 1 miljon sõna, mis koosneb järgmistest tekstidest.

| **Valdkond** | **Sõnade arv** | **Protsent korpusest** |
|---|---|---|
| Ajakirjandus | 175 000 | 17,5 % |
| Dokumendid | 12 000 | 1,2 % |
| Entsüklopeedilised teosed | 20 000 | 2,0 % |
| Esseed ja biograafiad | 90 000 | 9,0 % |
| Hobid ja harrastused | 75 000 | 7,5 % |
| Ilukirjandus | 250 000 | 25,0 % |
| Populaarteadus | 150 000 | 15,0 % |
| Propaganda | 60 000 | 6,0 % |
| Religioon | 8 000 | 0,8 % |
| Teadus | 160 000 | 16,0 % |

Tabel 2 - Tekstiklassid tänapäeva eesti kirjakeele baaskorpuses[9]

Baaskorpuses on esialgu ühestatud substantiive ja verbe (Kerner 2007, 10).

### 1.3.1. Sõnatähenduse ühestamise analüüsi etapid

Käesoleva peatüki kirjutamiseks on kasutatud materjali (Müürsep).

Sõnatähenduse ühestamise juures on esimesena asjana vaja teada saada, millisesse sõnaliiki antud sõna kuulub. Inglise keele puhul morfoloogiline analüüs ja ühestamine enam-vähem sellega piirdubki. Ingliskeelsetest lausetes ’He did his first move(n)’ ja ’He planned to move(v)’ on pärast morfoloogilist ühestamist kohe ka sõna tähendus selge.

Sõnatähenduse ühestamine on jagatud erinevateks etappideks. Neist olulisemad on järgmised etapid:

1. Eeltöötlus, mille käigus tuntakse ära lause lõpud ja tehakse kirjavahemärkide analüüs. Tekst teisendatakse morfoloogiaanalüsaatori jaoks sobivale kujule.
2. Morfoloogiline analüüs, mille käigus leitakse iga sisendsõna jaoks sõnavormi tüve ning lõpud ja neile vastava sõnaliigi, käände või pöörde. Kui sõnavorm on mitmeti tõlgendatav, antakse sellele kõik tõlgendused.
3. Morfoloogiline ühestamine, mille käigus valitakse morfoloogiliselt analüüsitud lause igale sõnale tema kõikvõimalike märgendite hulgast korrektne tõlge.
4. Sõnatähenduse ühestamine, kus sõnale leitakse sobiv tähendus.

[9] http://www.cl.ut.ee/korpused/baaskorpus/1980_stat (21.03.2013)

Paljud tekstid sisaldavad pealkirju, lõikude nummerdamist, suur- ja väiketähti. Selleks, et neid tekste analüüsida, tuleb nad viia mingile kindlale normaalkujule. Eeltöötluse käigus tuleb kindlaks määrata lausete piirid. Järgmisel analüüsietapil tehakse kõikidele sõnadele morfoloogiline analüüs. Morfoloogiaanalüsaator annab iga sisendsõna jaoks tema kõik morfosüntaktilised interpretatsioonid.

```
Aknas
    aken+s //_S_ com sg in //
kustus
    kustu+s //_V_ main indic impf ps3 sg ps af //
tuli
    tule+i //_V_ main indic impf ps3 sg ps af //
    tuli+0 //_S_ com sg nom //
$.
    . //_Z_ Fst //
```

Joonis 1 – Morfoloogiliselt analüüsitud lause

Morfoloogilise analüüsi käigus ei arvestata sõna konteksti. Morfoloogilise ühestamise etapil aga püütakse eemaldada kõik sellised tõlgendused, mis antud konteksti ei sobi. Selleks kasutatakse morfoloogilisi kitsendusi, mis leiavad lause igale sõnale korrektse morfoloogilise tõlgenduse. Joonisel 1 on kujutatud morfoloogiliselt analüüsitud teksti, kus sõnale *tuli* on leitud kaks tähendust. Õige tähenduse leidmiseks kasutatakse kitsenduste rakendamist – eemaldada tuleb verbi pöördeline vorm, kui antud sõnale eelneb vahetult verbi pöördeline vorm. Morfoloogiliselt ühestatud lause on Joonisel 2.

```
Aknas
    aken+s //_S_ com sg in //
kustus
    kustu+s //_V_ main indic impf ps3 sg ps af //
tuli
    tuli+0 //_S_ com sg nom //
$.
    . //_Z_ Fst //
```

Joonis 2 - Morfoloogiliselt ühestatud lause

Morfoloogiline analüsaator eristab üheksat sõnaliiki, milleks on substantiiv, adjektiiv, pronoomen ja numeraal kui käändsõnad, verb kui pöördsõna ja adverb, pre- ja postpositsioon, konjunktsioon ning interjektsioon kui muutumatud sõnad. Verbid omakorda on jagatud põhi-, abi- ja modaalverbideks; nimisõnad jagunevad üld- ja pärisnimedeks, arvsõnad põhi- ja järgarvsõnadeks, kaassõnad ees- ja tagasõnadeks. Asesõnu liigitati varsemal morfoloogilisel ühestamisel kaheksasse alaliiki, kuid sellest loobuti. (Vider, Muiscnek 2004)

### 1.3.2. Käsitsi sõnatähenduste ühestamine

Eesti keele puhul on käsitsi ühestamise aluseks võetud *wordnet*-tüüpi leksikon TEKsaurus; tekstis olevaid substantiive, põhiverbe ja modaalverbe ühestatakse TEKsauruse tähendusnumbrite põhjal (Kerner 2007, 11).

Enne kui saame teha semantilist analüüsi on vaja teha morfoloogiline analüüs, mille teeb eesti keele jaoks ESTMORF[10].

Ühe teksti ühestamisega tegelevad kaks inimest, sobivaim tähendusenumber kirjutatakse sõna reale. Seejärel ühestatakse erinevalt ühestatud sõnade tähendused. Kui alusleksikonist ühest tähendust ei leita, võibki sõna jääda mitmeseks. Tekstid, mida ühestatakse, on võetud Eesti Kirjakeele korpusest[11]. (Lõpparuanne 2003-2006)

Sõnavormi õige lemma taha märgitakse plussmärgiga sõna tunnus. Kui ühtegi käändelõppu/pöördetunnust ei lisandu, või on tegu muutumatu sõnaga, siis lisatakse plussmärgi taha 0. „// //“ märkide vahele lisatakse morfoloogiline info. Morfoloogilise info järele tuleb käsitsiühestajate poolt märgendatud sõnatähendus. Kui alusleksikonist sobivat märgendit ei leita, lisab *Semyhe* +1. @ märgi järele lisatakse sõna semantiline info, kaasa arvatud TEKsauruses esinev tähendusenumber. (Kerner 2007, 11)

[10] http://www.eki.ee/keeletehnoloogia/projektid/estmorf/ (21.03.2013)
[11] http://www.cl.ut.ee/korpused/baaskorpus/ (21.03.2013)

```
Ülejärgmisel
    ülejärgmine+l // _A_ pos sg ad //
päeval
    päeval+l // _S_ com sg ad // 6 @ päev:1711:8
oli
    ole+i // _V_ aux indic impf ps3 sg ps af //
õnn
    õnn+0 // _S_ com sg nom // 1 @ õnn:492#8698#8938:3
```

Joonis 3 - Näide käsitsi ühestatud failist.

## *1.4. Automaatne sõnatähenduste ühestamine*

Automaatse sõnatähenduse ühestamiseks on vajalik tekstis teatud sõnad enne käsitsi ühestada, luua semantiliselt ühtlustatud treeningkorpus ja seejärel rakendada sama teksti samadele sõnadele vastavat automaatse ühestamise meetodit. (Talve 2005, 14)

„Automaatse sõnatähenduste ühestamise eesmärgiks on käsitleda tekstis leiduvaid sõnu kui erinevate tähenduste hulki ning fikseerida iga hulga puhul üks tähendus, mis on antud kontekstis õige.“ (Talve 2005, 14) Seoses sellega tuleb esimesena kindlaks teha vaadeldava sõna kõik tähendused. Selleks võib kasutada erinevaid sõnastikke (tesauruseid, elektroonilisi- või tõlkesõnastikke). Seejärel tuleb teha kindlaks, milline tähendus on antud kontekstis kõige sobivam. (Talve 2005, 14)

Praeguseks on eesti keelele loodud automaatne morfoloogiline analüüs ja süntees ning süntaktiline analüüs, kuid täielik automaatne semantiline analüüs veel puudub.

### 1.4.1. Eesti keele automaatne sõnatähenduste ühestaja – Semyhe

Automaatseks ühestamiseks kasutatakse Tartu Ülikooli arvutilingvistika uurimisgrupis Eesti Kirjakeele korpuse tekste ja Kaarel Kaljuranna poolt loodud programmi *Semyhe,* mis kasutab Wordneti hierarhiaid ja on mõeldud ühestama substantiive ja verbe.

Programm *Semyhe* on koostatud programmeerimiskeeles Perl. Sisendtekst peab olema morfoloogiliselt analüüsitud. Programmi töö käigus lisatakse väljundile semantiline analüüs. *Semyhe* leiab sõna tähenduse TEKsaurusest. Viimane versioon programmist peaks leidma igale morfoloogiliselt ühestatud sõnale ühe vaste. Teksti morfoloogiline analüüs

tehakse ESTMORF programmi poolt. Viimane versioon programmist peaks leidma igale morfoloogiliselt ühestatud sõnale ühe vaste. Teksti morfoloogiline analüüs tehakse ESTMORF programmi poolt. *Semyhe* väljundis märgitakse morfoloogilise informatsiooni kõrvale ka semantiline kirjeldus – tähendusenumbrid. (Kerner 2004, 20)

## *1.5. Ühestamise meetodid eesti keele jaoks*

Sõnatähenduse ühestamiseks on mitmeid meetodeid, mida saab jagada kahte peamisesse kategooriasse. Esimene selline meetod põhineb grammatikareeglitel. Reeglipõhise meetodi puhul koostatakse reeglite komplekt sõnade järgnevuse alusel sõnavormi määramiseks. Tihtipeale on selliste reeglite loomine keeruline protsess. Teine meetod on statistiline. Selle lähenemisviisi jaoks arvutatakse sõna tõenäosuse osakaal antud kontekstis. Kõrgeima tõenäosusega tähendus omistataksegi antud sõnale. (Kerner, Vider, Neeme, 2006) Vähem on kasutatud ühestamisel neurovõrke.

Inimeste omavahelises suhtlemises sõnatähenduse ühestamine endast probleemi ei kujuta, sest õige tähenduse valimist toetab eelkõige kontekst, nii keeleline kui pragmaatiline (Langemets).

### 1.5.1. Reeglipõhine ühestamine

Reeglipõhise ühestamise puhul koostatakse reeglite komplekt, mis on aluseks reeglite formaliseerimiseks arvutile. Keele grammatika koostamiseks ei saa loendada lauseid, vaid tuleb panna kirja keele reeglid. Reeglipõhist ühestamist kutsutakse ka ratsionalismiks. Reeglipõhiste mudelite eelisteks on see, et reeglid on pööratavad (rakendatavad nii analüüsiks kui ka sünteesiks) ning suudavad efektiivsemalt käsitleda kaugsõltuvusi (öeldise ja aluse ühildumine) kui statistilised mudelid. Samuti on nendel mudelitel ka puuduseid. Nad on tundlikumad sisendi väikesemate kõrvalekallete suhteski ning reeglite väljatöötamiseks on vaja häid asjatundjaid, kuna need mudelid ei suuda näidetest õppida. [12] Vaata ka peatükki 2.1 Reeglid ja nende formaliseerimine.

### 1.5.2. Statistiline ühestamine

Statistilist ühestamist kutsutakse empirismiks. Statistiliste keelemudelite eelisteks on efektiivne tüüpilise keelekasutuse käsitlemine, juhul kui neid on korpusel treenitud, ja nad ületavad reeglipõhiseid mudeleid näiteks kõne puhul. Ka need mudelid ei ole puudusteta.

[12] http://zzz.ee/index.php/Ratsionalism_ja_empirism (10.05.2013)

Selle keelemudeli miinuseks on treenimiseks vajalike korpuste kogumine ja töömahukas ning veaohtlik märgendamine.[13]

Statistilise ühestamise puhul on enamlevinud meetodiks Markovi peitmudel (ingl k *Hidden Markov Model* – HMM).

Statistilise ühestamismeetodi korral konstrueeritakse alguses sõnaliikide esinemise tõenäosuste tabelid. Osad meetodid vajavad eelnevalt käsitsi ühestatud tekste, kuid on ka meetodeid, mis konstrueerivad tabeleid ühestamata tekstidest lähtudes, iteratiivselt treeningtekste ühestades ja nende põhjal tabeleid koostades. Käsitsi ühestatud tekstide kasutamine annab üldjuhul paremaid tulemusi. Statistilisel ühestamisel on märgendite valik väga oluline, see eristab head ühestajat halvast. Samas ei ole olemas häid eeskirju märgendite süsteemi tegemiseks. (Kaalep)

### 1.5.3. Hetkeolukord

Hetkeseisuga on eestikeelsete morfoloogiliselt ühestatud tekstide olukord selline, et ca 500 000 sõnaga korpust on teineteisest sõltumatult ühestanud vähemalt kaks inimest; kolmas on tulemused hiljem ühtlustanud.

Kõik ilukirjanduse tekstid on pärit eesti autorite töödest. Ajakirjanduse tekstid kuuluvad vahemikku 1995-1999 ja on võetud lehtedest „Äripäev“, „Eesti Ekspress“, „Eesti Päevaleht“, „Maaleht“, „Postimees“ ja „Sõnumileht“. Seaduse tekstid on võetud ÕTK koduleheküljelt www.legalext.ee aprill 2002 seisuga ja ka mujalt. Artikleid on võetud ka horisondi koduleheküljelt www.horisont.ee oktoober 2003 seisuga.

Tekstid kuuluvad järgmistesse klassidesse (sõnade hulka ei ole arvestatud kirjavahemärke):

| **Liik** | **Sõnade arv** |
|---|---|
| Ilukirjandus (Eesti autorid) | 104 000 |
| G. Orwelli „1984“ | 75 500 |
| Ajakirjandus | 111 000 |
| Seadused | 121 000 |
| Horisont | 98 000 |
| Info-tekstid | 4 000 |
| Kokku | 513 000 |

Joonis 2 – Morfoloogiliselt ühestatud korpuse tekstide jaotus

[13] http://zzz.ee/index.php/Ratsionalism_ja_empirism (10.05.2013)

Morfoloogia-analüsaator eristab 17 erinevat sõnaliigi märgendit.
Kasutatakse järgmisi sõnaliigi märgendeid (Kajaste 2009, 26):

- _A_ omadussõna - algvõrre (adjektiiv - positiiv), nii käänduvad kui käändumatud, nt *kallis* või *eht*
- _C_ omadussõna - keskvõrre (adjektiiv - komparatiiv), nt *laiem*
- _D_ määrsõna (adverb), nt *kõrvuti*
- _G_ genitiivatribuut (käändumatu omadussõna), nt *balti*
- _H_ pärisnimi, nt *Kristjan*
- _I_ hüüdsõna (interjektsioon), nt *tere*
- _J_ sidesõna (konjunktsioon), nt *ja*
- _K_ kaassõna (pre / postpositsioon), nt *kaudu*
- _N_ põhiarvsõna (kardinaalnumeraal), nt *kaks*
- _O_ järgarvsõna (ordinaalnumeraal), nt *teine*
- _P_ asesõna (pronoomen), nt *see*
- _S_ nimisõna (substantiiv), nt *asi*
- _U_ omadussõna - ülivõrre (adjektiiv - superlatiiv), nt *pikim*
- _V_ tegusõna (verb), nt *lugema*
- _X_ verbi juurde kuuluv sõna, millel eraldi sõnaliigi tähistus puudub, nt *plehku*
- _Y_ lühend, nt *USA*
- _Z_ lausemärk, nt -, /, ...
- _T_ tundmatu sõna

# 2. Programmi ülevaade

Lõputöös valminud programmi (regyhe) eesmärgiks on etteantud reeglite põhjal sõnatähenduste automaatne ühestamine. Sõnatähenduste ühestamise eelduseks on morfoloogiliselt analüüsitud ja ühestatud tekst.

Sõnatähenduse ühestamisel kasutan reeglipõhist ühestamise meetodit, mis on suhteliselt vähelevinud, sest selle meetodi puhul moodustatakse inimesele arusaadavad keelereeglid, mida järk järgu haaval rakendades välistatakse valed märgendid. Tuntuim selline reeglipõhise ühestamise meetod on kitsenduste grammatika. (Kajaste 2009,7)

Sõnatähenduste ühestamise reeglid on kirja pandud praegu eestikeelsete lausetena, mis on abiks leksikograafidele õige sõnatähenduse määramisel. Hetkel on reegleid umbes 100 ja need kirjeldavad sagedamini esinevaid polüseemseid nimisõnu.

Programm on loodud programmeerimiskeeles Python ja ühildub versioonidega 2.4 kuni 2.7.4.

## *2.1. Reeglid ja nende formaliseerimine*

Käesolevas töös on kasutatud Kadri Vare poolt kirjutatud reegleid[14] ning need on loodud selleks, et aidata leksikograafe, kes märgivad tähendust. Näide ühest sellisest reeglist:

JUHIS: vali hing(4):

kui sõna hing esineb ainsuse illatiivis

Teiseks ei mahtunud talle hästi hinge, et ... (tkt0111)

Hinge poeb vimm (tkt00110)

kui sõna hing esineb ainsuse inessiivis

...nagu peituks kunstniku hinges midagi erakordselt (tkt0111)

...oli Pärtli hinges iselaadi ja seletamatu rahutus (tkt0020)

Reeglite formaliseerimisel eristasin järgmisi sümboleid:

$ - tähistab morfoloogilise info algust

% - tähistab lemma algust

. - tähistab eelnevat sõna

/ - tähistab järgnevat sõna

— - tähistab eitust

[14] Substantiivide ja verbide reeglid http://www.keeletehnoloogia.ee/ekt-projektid/semantika-vahendid-eesti-keelele/reeglid (04.05.2013)

Lisaks on kasutatud loogikaavaldisi - konjunktsiooni ja disjunktsiooni.
Vastav formaliseeritud reegel ülaltoodule on: hing 4 -> [$_S_ com sg ill] v [$_S_ com sg in]

## *2.2. Kasutatavad tehnoloogiad ja algoritmid*

Selleks, et kasutada reeglipõhist ühestajat, on vaja tekstid kõigepealt töödelda sobivale kujule. Selleks vajab programm korpust, milles iga sõna on eraldi real, sõnale järgnevad tühikud ja seejärel on sõna märgend.

Reeglite formaliseerimisel on kasutatud Dijkstra Shunting-yard algoritmi[15].

## *2.3. Kasutusjuhend*

### 2.3.1. Programmi käivitamine

Selleks, et käivitada programm, tuleb:

1. Kopeerida failid regyhe kaustast oma arvutisse.
2. Veenduda, et arvutisse on installeeritud python versioon 2.4 - 2.7.4.
3. Avada käsurida ja liikuda kausta, kuhu failid kopeeriti.
4. Programmi käivitamiseks tuleb käsurealt ette anda 3 muutujat, vastavalt siis peaklass, sisend- ja väljundfail. Näiteks: python regyhe.py input.txt output.txt

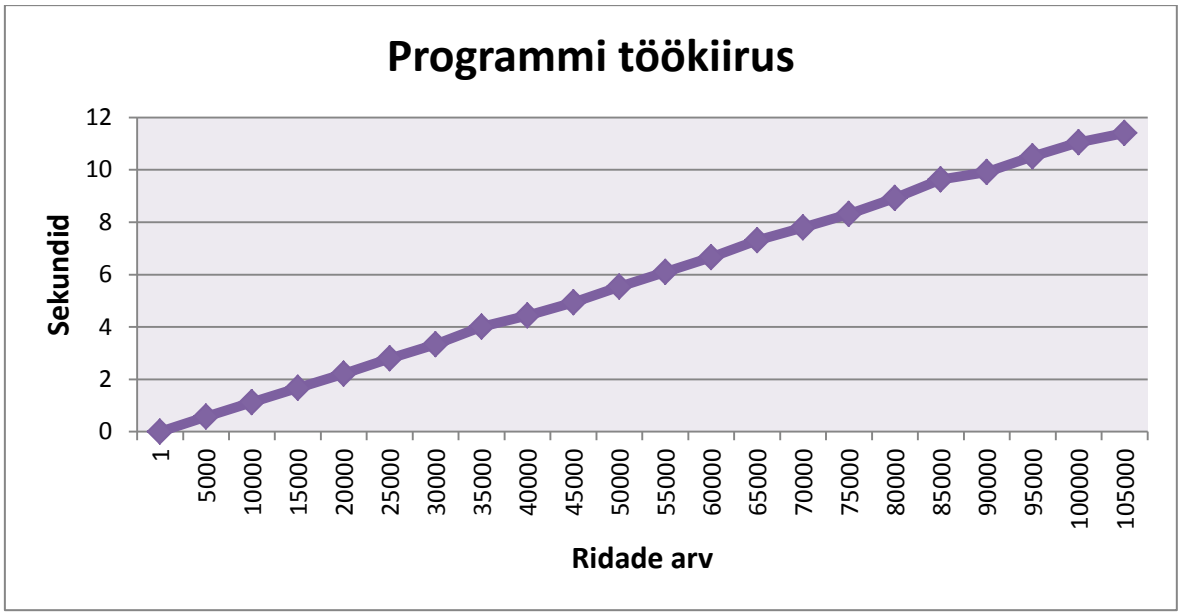
### 2.3.2. Programmi sisend ja väljund

Programmi sisendfaililt eeldatakse UTF-8 kodeeringut. Programmi sisend- ja väljundfaili laienditeks soovitatakse kasutada .txt või .kym laiendeid. Programm ei tööta .docx laiendiga.

## *2.4. Programmi testimine ja töökiirus*

Programmi on testitud erinevate morfoloogiliselt ühestatud tekstidega, mis on pärit morfoloogiliselt ühestatud korpusest[16]. Samuti on uuritud programmi töökiirust, mida illustreerib Joonis 4. Testimisel kasutati Windows keskkonda ja Pythoni versiooni 2.7.4.

[15] http://en.wikipedia.org/wiki/Shunting-yard_algorithm (10.05.2013)
[16] http://www.cl.ut.ee/korpused/morfkorpus/myh01/ (04.05.2013)




Joonis 4 - Programmi töökiirus

# Kokkuvõte

Sõnatähenduste ühestamine on semantilise ühestamise üks allülesandeid. Selle käigus omistatakse sõnale just see tähendus, mis tuleneb tema kontekstist. Erinevates kontekstides võib ühel sõnal olla erinevad semantilised interpretatsioonid, milleks on homonüümia ja polüseemia. Sõnatähenduse ühestamine käib mingi etaloni alusel, milleks on eesti keeles TEKsaurus. Tema väikseim osa on sünohulk ehk sünonüümirida, mille moodustavad ühte mõistet väljendavad sünonüümsed (sama tähendusega) sõnad ja sõnaühendid.

Ühestamisel on kasutusel mitmed meetodid, peamiselt kasutatakse reeglipõhiseid ja statistikal põhinevaid ühestajaid. Käesoleva töö teoreetilises osas antakse ülevaade sõnatähenduse ühestamise erinevatest mudelitest ja käsitsi ning automaatse ühestamise meetoditest. Hetkel on eesti keele jaoks olemas umbes 500 000 sõnast koosnev morfoloogiliselt ühestatud korpus, mida on ühestanud vähemalt kaks inimest.

Praktilise osa eesmärgiks oli formaliseerida olemasolevad sõnatähenduste ühestamise reeglid ja luua programm, mis kasutaks neid reegleid sõnatähenduste märgendamiseks korpuses. Töö käigus formaliseeriti 75 nimisõna ja 5 verbi reeglit. Sõnatähenduste ühestamise reeglid olid seni kirja pandud eestikeelsete lausetena, mis olid abiks leksikograafidele õige sõnatähenduse määramisel.

**A Rule-Based Disambiguator for Estonian**

Bachelor's thesis

Kristi Zirk

# Summary

Word-sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word is used in a sentence, when the word has multiple meanings. WSD is performed by using TEKsaurus as a reference sense inventory for Estonian. The atom of a wordnet-type thesaurus is a synonym set (also called a synset), which is a set containing all the synonymous words or multi-word units that express the same concept.

WSD can be classified into two categories: rule-based method and statistics-based method. The theoretical part gives an overview of general topics in WSD. Theoretical part also shows the process of manual and automatically WSD. At this moment morphologically disambiguated corpus of Estonian texts consists approximately 500 000 words and at least two people have disambiguation this.

The aim of the practical part was to formalize existing word-sense disambiguation rules and create a program what use these formalized rules to tag words in corpus. 75 noun and 5 verb rules were formalized during the work. WSD rules were so far written down in the Estonian sentences what were helpful to lexicographer to determining the proper meaning of the word.

# Kirjandus


Agirre, Eneko; Edmonds, Philik. Introduction 2006. – Word Sense Disambiguation. Ed E.Agirre, P. Edmonds. Netherlands: Springer, Seris: Text, Speech and Language Technology, Vol.33.

Ide, Nancy; Veronis, Jean (1998). Introduction to the special isuse on word senise disambiguation: the state of the art. – Computational Linguistics, No 24.

Kaalep, Heiki-Jaan. Analüüsivariantide hulgast valimine konteksti alusel. http://kodu.ut.ee/~hkaalep/arvutimorf_09/loeng7.htm (07.04.2013)

Kajaste, Kadri (2009). Eestikeelsete tekstide morfoloogiline ühestamine. Bakalaureusetöö.

Kerner, Kadri (2004). Sõnatähendused tekstides ja tesauruses ühestajate erimeelsuste põhjal. Bakalaureusetöö.

Kerner, Kadri; Vider, Kadri; Kahusk Neeme (2006). Sõnatähendused ja nende ühestamine tekstides. - *Keel ja arvuti. TÜ üldkeeleteaduse õppetooli toimetised 6*, lk 97-104

Kerner, Kadri (2007). Sõnatähenduste ühestamise tulemuste parandamise meetodeid eesti keele jaoks. Magistritöö.

Langemets, Margit. Polüseemia ja leksikograafia. http://www.emakeeleselts.ee/esa/ESA_49_pdf/Langemets.pdf (10.05.2013)

Lõo, Kaidi (2010). Püsiühendid ja liitsõnad *wordnet*-tüüpi tesauruses. Bakalaureusetöö

Müürsep, Kaili. Analüüsi etapid. http://www.cs.ut.ee/~kaili/parser/demo/ (02.05.2013)

Orav, Heili (2011). Eesti Wordnet'i täiendamine. http://www.keeletehnoloogia.ee/ekt-projektid/eesti-wordneti-taiendamine (18.03.2013)

Talve, Birgit (2005). Tekstide kaudu tuvastatud eesti keele tesaurusest puuduvad sõnatähendused. Bakalaureusetöö.

Tähenduspõhise keeletöötluse ressursid ja töövahendid eesti keele jaoks. 2003 – 2006. ETF grant nr 5534 (2003-2006) lõpparuanne (käsikiri). Tartu.

Vider, Kadri; Muiscnek, Kadri (2004). Sõnaliigituse kitsaskohad eesti keele arvutianalüüsis. Eesti Rakenduslingvistika ühingu aastaraamatu 1(2004). Eesti Keele Sihtasutus, Tallinn 2005, lk 99-114.

Õim, Katre (2012). Tähenduse varieerumine: polüseemia, homonüümia ja ebamäärane tähendus http://www.tlu.ee/~jaanike/loengud/T%C3%A4henduse_varieerumine.pdf (02.05.2013)

# Lisad

# Lisa 1. CD lähtekoodiga

CD sisaldab programmi lähtekoodi ja näidissisend faile. Tabelis 3 on toodud CD-l sisalduvad failid.

| **Failinimi** |
|---|
| Input.txt |
| Input2.txt |
| Logictree.py |
| Main.py |
| Parse.py |
| README.txt |
| Regyhe.py |
| Rulereader.py |
| Rules.txt |
| Termparser.py |

Tabel 3 - Nimekiri CD-l sisalduvatest failidest

**Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks**

Mina _____ Kristi Zirk _____

(*autori nimi*)

(sünnikuupäev: _____19.11.1990_____)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose
Reeglipõhine ühestaja eesti keele jaoks
_____
_____,
(*lõputöö pealkiri*)

mille juhendaja on _____ Neeme Kahusk_____
(*juhendaja nimi*)

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **13.05.2013**