

TARTU ÜLIKOOL
FILOSOOFIATEADUSKOND
Eesti ja üldkeeleteaduse instituut

Kadri Kerner

Sõnatähenduste ühestamise tulemuste parandamise meetodeid eesti keele jaoks
Magistritöö

Juhendaja Kadri Vider, M.A.

Tartu 2007

Sisukord

Sissejuhatus.....	3
1. Sõnatähenduste ühestamine	6
1.1. Sõnatähenduste ühestamine keeletehnoloogilistes rakendustes.....	6
1.2. Sõnatähenduste ühestamise alusleksikon.....	7
1.2.1. TEKsaurus	8
1.3. Sõnatähenduste ühestatud korpused.....	9
1.3.1. Sõnatähenduste ühestatud korpused maailmas	9
1.3.2. Eesti sõnatähenduste ühestatud korpus	10
1.3.3. Käsitsi sõnatähenduste ühestamine	11
1.4. Automaatne sõnatähenduste ühestamine.....	12
1.4.1. Automaatse sõnatähenduste ühestamise meetodid.....	12
1.4.2. Automaatsete sõnatähenduste ühestamissüsteemide hindamine – Senseval	14
1.4.3. Eesti keele automaatne sõnatähenduste ühestaja – Semyhe	15
2. Automaatse sõnatähenduste ühestamise ülesande tulemuste parandamise meetodeid. 17	17
2.1. <i>wordnet</i> -tüüpi leksikonides olevate sarnaste sõnatähenduste klasterdamine.....	17
2.1.1 Semantilise peegeldamise meetod	19
2.1.2. Regulaarne polüseemia klasterdamisel	22
2.1.3. Käsitsiühendajate eriarvamuste põhjal klasterdamine.....	24
2.1.4. Semantiliste suhete abil klasterdamine	25
2.2. Võimalikud täiendused automaatsele sõnatähenduste ühestamise süsteemile.....	26
2.2.1. Printsip „üks tähendus ühe teksti kohta“	26
2.2.2. Kontekstist tulenevad reeglid, morfoloogiline info	27
2.2.2.1. Grammatiseerunud vormid.....	27
3. Materjali valik	29
4. Märendusjuhiste loomine	30
4.1. Juhised.....	31
5. Klasterdamine	75
5.1. „Semantilise peegeldamise“ meetodi kasutamine klasterdamisel.....	75
5.2. Regulaarse polüseemia põhimõtete kasutamine klasterdamisel	89
6. Tulemused, järeldused	92
7. Kokkuvõte.....	98
Kirjandus	99
Methods of Improving Word Sense Disambiguation for Estonian Language. Summary	104
Lisa 1 – Analüüsis viidatud käsitsiühendajate erimeelsuste failid.....	105
Lisa 2 – Analüüsitud substantiivide esinemissagedused STÜ korpuses.....	107
Lisa 3 – CD-ROM. Printsipi „üks tähendus ühe teksti kohta“ järgivate sõnade loend; printsipi „üks tähendus ühe teksti kohta“ mitte järgivate sõnade loend	

Sissejuhatus

Sõnatähenduste ühestamine (STÜ, ingl *word sense disambiguation*) on üks semantilise analüüsi allülesandeid, seega on ta ka üks olulistest ülesannetest, mis tuleb lahendada selleks, et keeletehnoloogilised rakendused tuleksid toime loomuliku keelega. STÜ on keeletehnoloogiline ülesanne, mille puhul omistatakse sõnale just see tähendus, mis tuleneb tema kontekstist (Agirre, Edmonds 2006: 2). Nii morfoloogiline kui süntaktiline analüüs on oma töös saavutanud juba peaaegu täiusliku täpsuse, kuid semantiline analüüs on veel siiski lahendamata probleem. Väidetakse isegi, et keele automaatses analüüsis ongi STÜ probleem olulisim (Ide, Veronis 1998: 2). Kuid siiani ei ole saavutatud automaatses sõnatähenduste ühestamises piisavat täpsust. Muuta STÜ ülesanne kergemini teostatavaks on ka käesoleva töö eesmärk.

Antud töö esimene, teoreetiline osa annab ülevaate sõnatähenduste ühestamisega seotust. Oluline on teada, missuguseid sõnu on kõige raskem ühestada – sõnad, millel on ainult üks tähendus, ei kujuta automaatsele ühestamissüsteemile probleemi, põhiline on just mitmetähenduslike sõnade ühestamine. Töös on kirjeldatud, millised keeletehnoloogilised rakendused vajavad ühestatud sõnatähendusi. Kuna maailmas on Princetoni WordNet võetud üheks peamiseks STÜ alusleksikoniks, siis on antud ülevaade nii sellest kui ka samadel põhimõtetel EuroWordNeti projekti raames loodud Eesti Wordnetist (ka: eesti keele tesaaurus ning TEKsaurus). Viimase põhjal toimub eesti keele puhul nii automaatne kui ka käsitsi STÜ. Ülevaade on antud sellest, mille jaoks ühestatakse sõnatähendusi käsitsi ning kuidas toimub käsitsi ühestamine Eestis. Käsitsi ühestatud sõnatähendused koondatakse ühestatud sõnatähenduste korpusesse (STÜ korpus) ning lühidalt on kirjeldatud olemasolevaid inglise keele kui ka eesti keele STÜ korpusi. Ka automaatne sõnatähenduste ühestaja võib oma töös rakendada just ühestatud korpusi, kuid ülevaade on antud teistestki võimalikest automaatsetest ühestamismeetoditest. Eraldi on kirjeldatud ka eesti keele automaatset sõnatähenduste ühestajat.

Teoreetilise osa teine peatükk keskendub sellele, missuguseid meetodeid on kasutatud, et muuta automaatne STÜ arvutile kergemaks ülesandeks. Välja on toodud sõnatähenduse definitsiooni problemaatika, s.o *wordnet*-tüüpi leksikonides olevad sõnatähendused on liiga granuleeritud ehk üle-eristatud. Siinkohal tuleb ära märkida, et üle-eristatuse all on silmas peetud mitte nähtust leksikonis, vaid pigem nähtust reaalses tekstikasutuses – on sõnu, mille tähendused ei tule tekstis välja või on neid tähendusi tekstis praktiliselt võimatu eristada. Liiga üle-eristatud sõnatähendustega alusleksikonile toetudes muutub ka automaatse ühestamissüsteemi töö keeruliseks. Erinevaid meetodeid kasutades on püütud sellist üle-eristatust vähendada, et saavutada optimaalne hulk sõnatähendusi erinevate keeletehnoloogiliste rakenduste jaoks. Peamine idee seisneb selles, et grupeerida tuleks väga sarnased sõna tähendused, ingliskeelses kirjanduses kasutatakse selle kohta terminit *cluster*, *clustering* ning ka käesolevas töös on kasutatud sõnu klaster ja klasterdamine just sarnaste sõnatähenduste grupeerimise mõttes.

Automaatse sõnatähenduste ühestamissüsteemi tööd saab hõlbustada erinevate lisakitsenduste näol ning nende võimalikest variantidest on antud põhjalikum ülevaade.

Käesoleva töö praktilises osas on analüüsitud mitmetähenduslikke nimisõnu. Analüüsi võib jaotada kolmeks osaks:

- 1) esiteks vaadatakse, milliste substantiivide puhul kehtib printsiip „üks tähendus ühe teksti kohta“;
- 2) siis analüüsitakse neid sõnu, mille puhul eelmainitud printsiip ei kehti, ning püütakse leida kontekstuaalseid reegleid või muid tingimusi, millisel juhul esineb mõni kindel sõnatähendus, ning luuakse märgendusjuhiseid;
- 3) kui sellised reegleid vms tingimusi ei leita, siis katsetatakse valitud sõnade puhul sarnaste sõnatähenduste klasterdamist.

Praktilise osa eesmärgiks on täiustada automaatset sõnatähenduste ühestamissüsteemi erinevate lisakitsendustega ning sobivama alusleksikoniga. Kuna töö on peamiselt

lingvistilise iseloomuga, siis on võimalik samu tähelepanekuid kasutada ka käsitsi ühestajate töö hõlbustamiseks.

Siinkohal tahan täna Kadri Viderit järjekindla juhendamise eest.

1. Sõnatähenduste ühestamine

STÜ on mingi keele semantilise analüüsi esimesi etappe. Eesti keeles algab sõnatähenduste ühestamine juba morfoloogilise analüüsi tasandil. Sisusõnade klassid, mille igal liikmel on lisaks kategoriaalsele tähendusele kindlasti ka iseseisev tähendus, on substantiivid, verbid, adjektiivid ja adverbid. Iseseisva tähendusega sõnadel on lisaks ka tendents mitmetähenduslikkusele, seega kasutatakse üht ja sama lekseemi mitmes tähenduses. Sisusõnale konteksti sobiva tähenduse leidmine ongi STÜ protsessi tulemus. (Lõpparuanne 2003-2006)

Sõnad, millel on mitu tähendust, on homonüümsed või polüseemsed. Homonüümid on samakujulised, aga erineva tähendusega sõnad, polüseemsed on need sõnad, mille kõik tähendused on omavahel mingit moodi seotud. Näiteks sõnad *pank*, *koer*, *tint* on homonüümid; sõnal *uks* aga on kaks omavahel tihedalt seotud tähendust – nii *takistus* kui *avaus* – ning seega on ta polüseemne sõna. Homonüümia on väidetavalt ajalooline või juhuslik; polüseemia võib tuleneda keele ökonoomsusest, kasutatakse ära juba keeles olemas olevad sõnad (Peters 2004: 16). Polüseemia võib tuleneda ka regulaarsest tähenduse laienemisest (ehk regulaarsest polüseemiast), seda võib põhjustada kontekst või võib polüseemia olla tingitud metafoorist ning metonüümiast (Ravin, Leacock 2002: 23). Mitmetähenduslikkus on keeles levinud, näiteks on nimisõnal *asi* eesti keeles (vastavalt TEKsaurusele) tervelt 12 erinevat tähendust, verbil *käima* 23 tähendust. Erinevate keeletehnoloogiliste rakenduste jaoks on vaja ühestada nii homonüümseid kui polüseemseid sõnu.

1.1. Sõnatähenduste ühestamine keeletehnoloogilistes rakendustes

STÜ on vahendava iseloomuga ülesanne, see võimaldab teistel rakendusel paremini toimida (Resnik 2006: 299).

Peamised loomuliku keele rakendused, mis vajavad oma töö täpsuse jaoks ka semantilist informatsiooni, seal hulgas ühestatud sõnatähendusi on esitatud alljärgnevas loetelus.

- Masintõlge - tõlkida tuleks õige tähendus. Kas sõna *naine* tõlkida inglise keelde kui *woman* või kui *wife* (abielunaine)?
- Info-otsing. Et otsida mõisteid ja tähendusi nt erinevates keeltes ning eemaldada need tähendused, mis pole otsingul olulised.
- Kõnetöötlus. Vaja on eristada homonüüme (just homograafe), näiteks *tulp:tulba* ja *tulp:tulbi*.
- Kõnedialoogsüsteemid, näiteks telefonisüsteemid, kus vastajaks ja suhtlejaks on arvuti.
- Loomulikus keeles suhtlust võimaldavad kasutajaliidesed. Näiteks ehitusealane kasutajaliides.
- Automaatne sisukokkuvõtete tegemine. (Ide, Veronis 1998: 1–2)

1.2. Sõnatähenduste ühestamise alusleksikon

Sõnatähendusi ühestatakse mingi etaloni alusel. Etalonis on eristatud sõnade (õigemini küll mõistete, mida sõnad keeles väljendavad) tähendused teatud tunnuste võrdlemise ja välistamise teel. Sel moel piiritletud sõnatähenduste kogum on STÜ alusleksikoniks, millest tekstisõna kontekstiga paremini sobivat tähendust otsitakse. (Lõpparuanne 2003–2006)

Princetoni WordNet¹i ning ka teiste keelte *wordnet*-tüüpi leksikaal-semantilisi andmebaase kasutatakse praeguse seisuga STÜ alusleksikonidena kõige enam. Selliste *wordnet*-tüüpi üheks põhiliseks rakenduseks nähakse just loomuliku keele rakendusi (Orav 1998: 15).

¹ <http://wordnet.princeton.edu/> (10.05.2007)

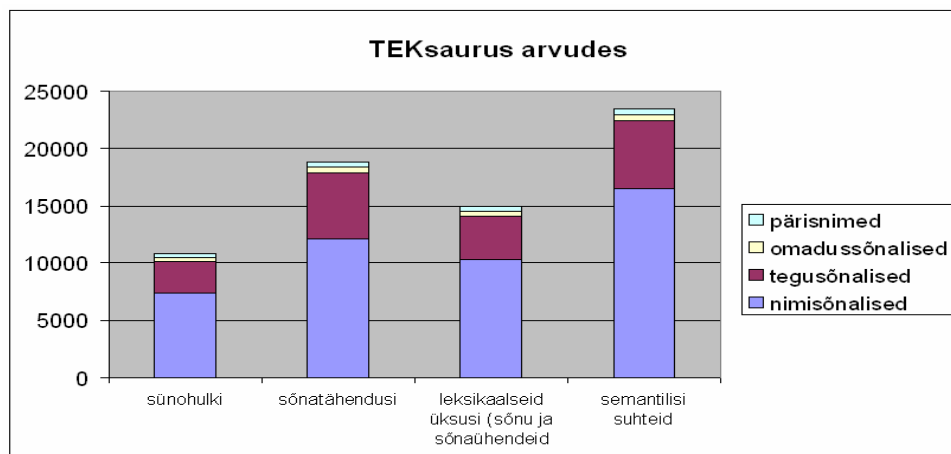
Eesti keele puhul on nii käsitsi kui ka automaatse STÜ jaoks alusleksikonina kasutatud eesti keele *wordnet*-tüüpi tesaaurust – TEKsaurust.

1.2.1. TEKsaurus

Eesti keele tesaauruse loomine käivitati 1998. aastal EuroWordNeti² projekti käigus, kus sarnaste põhimõtete järgi koostati kaheksa erineva keele tesaaurused. Kuna EuroWordNet on mitmekeelne, siis on osa TEKsauruses olevatest sünonühkadest³ ühendatud läbi ingliskeelse InterLingual Indexi (ILI-lingi ehk keeltevälise indeksi kaudu) ka teiste keelte tesaauruste sünonühkadega.

Eesti keele *wordnet*-tüüpi tesaauruses⁴ on sõnad koondatud mõistete alla nende tähenduse järgi, mis tähendab, et mitmetähenduslik sõna esineb nii mitmes mõistes, kui tal on tähendusi. Tähendusüksusi seovad omavahel 45 erinevat semantilist seost, nagu alam- ja ülemmõisted, antonüümia, osa-terviku suhe jms⁵. TEKsauruses on esindatud verbid, substantiivid, adjektiivid ning pärisnimed. Praeguse seisuga on TEKsauruses umbes 16 000 mõistet. (Lõpparuanne 2003–2006)

Tabel 1. TEKsauruse andmed⁶



² <http://www.illc.uva.nl/EuroWordNet/> (10.05.2007)

³ TEKsauruse elementaarosake on sünonüümirida; (ingl. *synonym set*, *synset*), mille moodustavad ühte mõistet (*concept*, *meaning*) väljendavad sünonüümsed (sama tähendusega) sõnad ja sõnaühendid.

⁴ <http://www.cl.ut.ee/ressursid/teksaurus/> (10.05.2007)

⁵ täielik loetelu vt <http://www.cl.ut.ee/ressursid/teksaurus/> (10.05.2007)

⁶ <http://www.cl.ut.ee/ressursid/teksaurus/> (10.05.2007)

1.3. Sõnatähenduste ühestatud korpused

STÜ korpus on tekstikorpus, milles ühes ja samas tähenduses sisusõnadel (lekseemidel) on viide ühele ja samale sõnatähendusele alusleksikonis. Inimeste annoteeritud ehk käsitsiühestatud STÜ korpus on vajalik nii sõnatähenduste piiride ja kasutuskonteksti uurimiseks konkreetses keeles kui ka mitmete keeletehnoloogiliste rakenduste treenimiseks või hindamiseks. Tekstide semantilises analüüsis peetakse käsitsiühestatud korpust etaloniks (nn. *Golden Standard*), milleni automaatselt ühestatud tekstid peavad oma täpsuses jõudma. (Lõpparuanne 2003–2006)

1.3.1. Sõnatähenduste ühestatud korpused maailmas

STÜ korpust on suhteliselt vähe, kuid põgus ülevaade on antud suurematest nii käsitsi märgendatud kui ka automaatselt märgendatud korpustest.

Käsitsi on ühestatud järgmised korpused:

- **DSO** (ingl *Defence Science Organisation*) **Korpus**⁷. See sisaldab tekste *Browni* ja *Wall Street Journali* korpustest. Ühestatud on inglise keele 121 nimisõna ja 70 verbi põhjal 192 800 sõnet. Alusleksikonina on kasutatud WordNeti 1.5 versiooni. (vt ka Ng, Lee 1997);
- **Semcor**⁸. Loodud samade inimeste poolt, kes koostasid Princetoni WordNeti. See on kõige mahukam STÜ korpus. Semcor koosneb 352st tekstist, millest 186 tekstis on ühestatud 192 639 sõnet (substantiivid, adjektiivid ja verbid) ning 166 tekstis ainult verbid, kokku 41 497 sõnet. Alusleksikonina on kasutatud WordNeti versiooni 1.6.
- **Open Mind Word Expert**. See korpus on märgendatud tavaliste (st mitteleksikograafide) inimeste poolt. Alusleksikonina on kasutatud WordNeti versiooni 1.7. (vt ka Mihalcea, Chklovski 2004)

⁷ <http://www ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97T12> (15.05.2007)

⁸ <http://multisemcor.itc.it/semcor.php> (15.05.2007)

Automaatselt ühestatud korpused on:

- **Sensecorpus**⁹. Korpus koosneb tuhandetest WordNeti versiooni 1.6 nimisõnade tähendustest.
- **ODP (ingl *Open Directory Project*) Corpus**¹⁰. Korpus koosneb WordNeti versiooni 1.7 nimisõnade tähendustest.

1.3.2. Eesti sõnatähenduste ühestatud korpus

Eesti STÜ korpus¹¹ koosneb kahest osast: baaskorpusest ja liikumislausete korpusest. Tekstid eesti STÜ korpuses on käsitsi märgendatud. Kuna antud töö praktilises osas on analüüsitud ühestatud sõnatähendusi just baaskorpuse põhjal, siis lähemalt antakse ülevaade just selle korpuse materjali ja andmete kohta.

Ühe sõna analüüsi tulemus on sarnane sellele, mida on kirjeldatud morfoloogiliselt ühestatud korpuse¹² juures, lisaks on info sõne tähenduse kohta ehk semantiline info:

sõne tüvi + lõpp // morfanalüüs // semantiline info

Baaskorpuses on 40 tekstilõiku 1980ndate aastate eesti ilukirjandust ja 1 tekstilõik ajakirjandust (1995. aasta Postimehest) – 113 870 tekstiüksust ehk sõnet, keskmiselt 2648 sõnet teksti kohta. Esialgu on ühestatud substantiive ja verbe. Tegu on nn. loomuliku keelekasutuse korpusega, mille laused pole konstrueeritud lingvistilise analüüsi eesmärkidel. (Lõpparuanne 2003–2006)

⁹ <http://ixa.si.ehu.es/Ixa/resources/sensecorpus> (10.05.2007)

¹⁰ <http://dmoz.org/> (10.05.2007)

<http://nlp.uned.es/ODP/> (10.05.2007)

¹¹ vaata ka: <http://www.cl.ut.ee/korpused/semkorpus/> (10.05.2007)

¹² www.cl.ut.ee/korpused/morfkorpus/ (10.05.2007)

1.3.3. Käsitsi sõnatähenduste ühestamine

Käsitsi ühestamise aluseks on võetud eesti keele puhul *wordnet*-tüüpi leksikon TEKsaurus; tekstis olevate verbide ja substantiivide tähendusi ühestatakse TEKsauruse tähendusnumbrite põhjal. Enne semantilise analüüsi tegemist on vaja teha morfoloogiline analüüs ning leida igale sõnavormile õige lemma ning morfoloogilise analüüsi teeb programm ESTMORF¹³. Ühte teksti ühestavad kaks inimest, sobivaim tähendusnumber kirjutatakse analüüsireale (vt joonis 1). Seejärel ühtlustavad käsitsimärgendajad erinevalt ühestatud sõnade tähendused. Võib juhtuda, et ka ühtlustamise käigus ei leita tekstisõnale ühte ja ainsat sobivat tähendust alusleksikonist, sellisel juhul võibki vastav tähendus mitmeks jääda. (Lõpparuanne 2003–2006)

```
Ülejärgmisel
ülejärgmine+l //_A_ pos sg ad //
päeval
päev+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 1. Näide käsitsi ühestatud failist.

Joonisel 1. on sõnavormi õige lemma taga on tunnused, märgitud plussmärgiga. Kui tunnust/käändelõppu/pöördetunnust ei lisandu või on tegu muutumatu sõnaga, siis lisatakse plussmärgi taha 0. '//' //' märkide vahel on morfoloogiline info. Pärast morfolooglist informatsiooni on käsitsiühestaja poolt märgendatud sõnatähendus; kui ühestaja ei leia alusleksikonist sobivat tähendust, lisab ta märgendi +1. @ märgile järgnev info on lisatud automaatselt programmi *semyhe* poolt. Viimase kooloni järel olev number näitab, mitu tähendust on antud sõnal TEKsauruses (Orav, Vider 2002: 300–301).

¹³ <http://www.eki.ee/keeletehnoloogia/projektid/estmorf/> (13.05.2007)

Käsitsiühestajatele on koostatud ka ühestamisjuhend, mis annab ühestamiseks vajalikke nõuandeid. See juhend on saanud järjest mahukamaks ja sisukamaks, ning täieneb edaspidi veelgi.

1.4. Automaatne sõnatähenduste ühestamine

Sõnatähenduste automaatse ühestamise vastu on huvi tuntud juba alates 1950ndatest aastatest, kui hakati tegelema masintõlkega. Vahepealne periood erilist edu ei toonud, kuid 1980ndatel, kui kättesaadavaks muutusid suuremahulised leksikaalsed ressursid, sai automaatne STÜ „uue hingamise“. 1990ndatest alates on sõnatähenduste ühestamisega tegeletud sageli just *wordnet*-tüüpi leksikonides olevate sõnatähenduste põhjal. (Ide, Veronis 1998: 3–13)

Automaatse STÜ probleemi aktuaalsust näitavad ka erinevate ühestamissüsteemide hindamiseks loodud organisatsioon SENSEVAL¹⁴ ning praeguseks läbi viidud juba kolm Sensevali konverentsi, millel on osalenud ka eesti keele sõnatähenduste ühestaja Semyhe (lähemalt vt ptk 1.4.3.).

1.4.1. Automaatse sõnatähenduste ühestamise meetodid

STÜ meetodid jaotatakse erinevatesse gruppidesse vastavalt ressursile, mida need ühestamise aluseks võtavad.

Neid meetodeid, mis toetuvad peamiselt sõnaraamatutele, tesaurustele ja leksikaalsetele teadmibaasidele, nimetatakse sõnaraamatul põhinevateks (ingl *dictionary-based*) või teadmistel põhinevateks (ingl *knowledge-based*) (Agirre, Edmonds 2006: 12). Kõige sagedamini kasutatakse alusleksikonina *wordnet*-tüüpi leksikoni. Peamiselt on kasutusel neli erinevat meetodit:

¹⁴ <http://www.senseval.org/> (10.05.2007)

- 1) Leski algoritm, mille puhul määratakse sõnale kõige tõenäolisem tähendus, mis esineb antud kontekstis (vt ka Lesk 1986);
- 2) meetodid, mis arvutavad semantilist sarnasust (ingl *semantic similarity*) semantilistes võrgustikes (vt ka Budanitsky, Hirst 2001, Leacock jt 1998);
- 3) nn valitud eelistuste meetod (ingl *selectional preferences*), mis kasutab tähenduste välistamist (vt ka Brockmann, Lapata 2003);
- 4) heuristilised meetodid, mis kasutavad printsiipe „üks tähendus ühe teksti kohta“ (vt ka ptk 2.2.1) ja „üks tähendus kollokatsiooni kohta“ ning kõige sagedamini esinevat tähendust (vt ka Gale jt 1992).

Juhendamata meetodid (ingl *unsupervised*) töötavad enamasti märgendamata korpustel ning kasutavad keelevälisest informatsiooni. Need meetodid ei kasuta alusleksikone ega STÜ korpusi (Pedersen 2006: 133–140). Meetodid toetuvad:

- 1) sõnade jaotusele (ingl *distribution*), mis ühestavad sõnatähendusi eeldusel, et sõnadel, mis esinevad sarnastes kontekstides, on sarnased tähendused (vt ka McCarthy jt 2004);
- 2) tõlkevastetele, mille abil luuakse lähtekeelele teatav sõnatähenduste komplekt võrreldes sihtkeeleaga (vt ka Resnik, Yarowsky 1997).

Juhendatud ja poolenisti juhendatud meetodil (ingl *supervised, semi-supervised*) ühestamine kasutab märgendatud korpuse põhjal treenitud süsteemi. Kasutatakse statistilisi ja reeglitel põhinevaid meetodeid. Tulemuslikuks on osutunud masinõppe meetodid ning loodud on palju erinevaid ühestamissüsteeme ning -algoritme, mida treenitakse STÜ korpusel. (Marquéz, Escudero, Martinez, Rigau 2006: 167–168)

1.4.2. Automaatsete sõnatähenduste ühestamissüsteemide hindamine – Senseval

Erinevatest ühestamismeetoditest ja tulemustest annavad hea ülevaate Sensevali konverentsid-võistlused. Esimene Sensevali võistlus toimus 1998. aastal ning sellel osales 24 erinevat ühestamissüsteemi nii juhendatud kui ka juhendamata meetoditega. Leksikaalne materjal, mille põhjal sõnatähendusi ühestati, pärines spetsiaalselt Sensevali jaoks loodud Hectori leksikonist, milles olid ühendatud nii korpusepõhised lähenemised kui ka sõnaraamatus olevad sõnatähendused. Paremad tulemused saavutasid üle 80%-lise täpsuse. (Palmer, Ng, Dang 2006: 87)

Teisel Sensevali konverentsil kasutati alusmaterjalina WordNetis olevaid sõnatähendusi. Seega muutus ühestamise ülesanne raskemaks, sest WordNetis on polüseemsete sõnade hulk suur ning polüseemsel sõnal on palju erinevaid tähendusi. Ülesanded jaotati laiemalt kaheks: valitud sõnade ühestamise ülesanne (ingl *lexical sample task*) ning kõikide sõnade ühestamise ülesanne (ingl *all-words task*). Mõlemate ülesannete puhul ning erinevaid meetodeid kasutades saavutati süsteemide täpsus vahemikus 65%–79%. (Palmer, Ng, Dang 2006: 88–89)

Senseval-3 toimus 2004. aastal ning seekord lahendasid 14 erinevat ülesannet 55 võistkonda 160 ühestamisüsteemiga. Ühestati nii WordNeti kui FrameNeti tähenduste põhjal, erinevate süsteemide täpsused jäid vahemikku 53%–85%. (Palmer, Ng, Dang 2006: 99–101)

Homonüümide eristamisega saavad automaatsed ühestamissüsteemid hakkama peaaegu 95%-lise täpsusega. Polüseemsete sõnade eristamine on palju keerulisem ülesanne, aja jooksul on küll meetodid ja ressursid paranenud, kuid ikkagi pole saavutatud täpsust üle 85%. Automaatse ühestamissüsteemi täpsuse tõstmiseks pakutakse välja erinevaid meetodeid – nii ressursside ümber tegemist kui ka ühestamissüsteemide täiustamist. (Agirre, Edmonds 2006: 15–16)

1.4.3. Eesti keele automaatne sõnatähenduste ühestaja – Semyhe

Sõna tähendus sõltub kontekstist. Ilma kontekstita ei ole põhimõtteliselt võimalik sõna tähendust määrata. Uusim versioon programmist Semyhe kasutab lokaalse konteksti puhul ka süntaktilise struktuuri kohta käivat informatsiooni, mida annab kitsenduste grammatika parser (ESTCG parser) (vt ka Müürisep 2000). Varasem versioon programmist Semyhe kasutas konteksti võrdlemiseks ainult TEKsauruses leiduvat informatsiooni. Nagu Semyhe eelminegi versioon, ei vaja ka praegune eelnevalt käsitsi ühestatud treeningkorpust. (Lõpparuanne 2003–2006)

Programmi Semyhe¹⁵ teostas Kaarel Kaljurand programmeerimiskeeles Perl. Ettevalmistaval etapil luuakse sõltuvusseoste andmebaas. Selles andmebaasis on iga süntaktilise struktuuri elemendi ja selle kontekstis leiduva verbi ja adjektiivide kohta vastavate nimisõnade sagedused, mida kasutatakse ühestamismudelite loomisel. (Lõpparuanne 2003–2006)

Kasutusel on kaks ühestamismudelit: M1 ja M2. Mudeli M1 puhul järjestatakse sarnased nimisõnad, mille sagedus samas kontekstis on suurem kui 1, vastavalt sellele kui mitmes analüüsitava sõna kontekstis nad esinevad, ning kui suur on nende poolt pakutud lähima tähenduse kaugus (WordNeti mõistes). Edetabeli kolm esimest sõna hääletavad, milline sõnatähendus valida. (Kaljurand 2004)

Mudel M2 toimib nagu M1, aga kontekstisõnadel, mis kuuluvad kümne sagedasema verbi/adjektiivi hulka, ei lasta “sarnaseid sõnu” genereerida, sest suur sagedus osutab kontekstisõna mitmetähenduslikkusele, mis toob kaasa väiksemaid semantilisi kitsendusi kontekstile. (Kaljurand 2004)

Tulemused, mis saadi püüdes määrata 1700 TEKsauruses kirjeldatud (nii ühetähendusliku kui ka mitmetähendusliku) nimisõna tähendust ilukirjandustekstides, on toodud järgmises tabelis :

¹⁵ http://www.psych.ut.ee/~kaarel/eng_semyhe/ (10.05.2007)

Tabel 2. Semyhe tulemused (Lõpparuanne 2003 – 2006)

Meetod	Täpsus	Saak
Juhuslik	44%	100%
Sagedaseim	82%	100%
Mudel 1	82%	62%
Mudel 2	89%	54%

Tabelis 2 esitatud meetod “juhuslik” omistab sõnale tähenduse juhusliku valiku teel. Meetod “sagedaseim” omistab sõnale selle kõige sagedasema tähenduse. Mudel1 ja mudel2 on eelpool kirjeldatud ühestamismudelid M1 ja M2. (Lõpparuanne 2003–2006)

2. Automaatse sõnatähenduste ühestamise ülesande tulemuste parandamise meetodeid

On jõutud arvamusele, et automaatset sõnatähenduste ühestamist on võimalik muuta hõlpsamini teostatavaks ülesandeks, kui teha ressursid sobivamaks ning täiustada automaatseid süsteeme. Paremate ning sobivamate ressursside all mõeldakse näiteks *wordnet*-tüüpi leksikonides olevate sarnaste sõnatähenduste klasterdamist erinevate meetodite abil ning ka spetsiifilise valdkonna¹⁶ või rakenduse jaoks vastava ressursi loomist või ümber kohandamist.

2.1. *wordnet*-tüüpi leksikonides olevate sarnaste sõnatähenduste klasterdamine

Väidetakse, et pole olemas halbu automaatseid ühestamissüsteeme, vaid ühestamise aluseks olevad ressursid on ebapiisavad (Navilgi 2006: 105). Kuna eesti keele puhul on valitud ühestamisel abistavaks ressursiks TEKsaurus, siis lähemalt tuleks vaadelda just sõnatähenduste eristatuse taset selles ning teisteski *wordnet*-tüüpi leksikonides .

Raske on määratleda, kus lõpeb ühe sõnatähenduse piir ning kust algab juba teine tähendus ning siiani on üheks suurimaks vaidlusküsimuseks see, kui eristatud sõnatähendused olema peavad (Kilgariff 1997: 92). Sõnatähendused ei tohiks loomulikult olla ei liiga üle-eristatud ega ka liiga ala-eristatud. Erinevad rakendused vajavad erinevat eristatuse taset – mis on piisav näiteks lihtsamale masintõlkesüsteemile, pole samas eriti informatiivne leksikograafidele. Ka keeletehnoloogilised rakendused vajavad erinevat eristatuse taset: info-otsing saab peaaegu 100%-se täpsuse ka väga üldise eristatusega, kuid masintõlkesüsteemile on vaja siiski pisut detailsemat sõnatähenduste eristatust. (Ide,

¹⁶ inglise keeles *domain*-, *discourse*-, *topic*-, *subject-specific*

Veronis 1998: 3)

Wordnet-tüüpi leksikonid on saanud leksikograafide töö tulemusena liiga üle-eristatud sõnatähendused. Sellised liiga peened sõnatähendused on aga üheks peamiseks probleemiks automaatse ühestaja jaoks, sest tekib palju müra ja palju vigu ning täpsus ja töökiirus vähenevad. Samuti muutub keeruliseks vajaliku STÜ korpuse käsitsi märgendamine, sest märgendajad ei suuda valida üht ja ainsat tähendust nendele sõnadele, millel on palju ja sarnaseid tähendusi. (McCarthy 2006: 17)

WordNeti tähenduste põhjal ühestatakse käsitsi umbes 70% tekstisõnu üksmeelselt, samasugune üksmeelsus saavutatakse ka TEKsauruse põhjal ühestades. Senseval-3 raames tehtud katsed näitavad, et polüseemiataseme mõistlikul määral vähendamisega on võimalik automaatse ühestamissüsteemi tööd oluliselt parandada. WordNetis olevate tähenduste arvu vähendati 33.54% (60 302-lt 40 079-le) ning polüseemiatase vähenes 3.14-lt 2.09-le. Kui enne oli polüseemsel sõnal keskmiselt 3.14 tähendust, siis pärast sarnaste sõnatähenduste grupeerimist vähenes see keskmiselt kahele tähendusele polüseemse sõna kohta. Kõige parema tulemuse sai vähemate tähendustega materjali põhjal ühestamissüsteem Gambl: 78% täpsust võrreldes üle-eristatud materjalil põhineva 65%-lise täpsusega. (Navigli 2006: 110)

Seega ongi aktuaalseks muutunud sarnaste sõnatähenduste klasterdamine.

Üle-eristatust saab vähendada erinevate klasterdamismeetoditega (ehk sarnaste tähenduste grupeerimisega), tavalisimad on:

- sagedus- ja korratusmaatriksi kasutamine (vt ka Chlkovsky, Mihalova 2003);
- sõnatähenduste näidete ja seletuste automaatne töötlemine, tõlkevastete kasutamine ning semantiline peegeldamine (vt ka Dyvik 2005; Ide, Erjavec, Tufis 2002; Saarikoski 2005);
- käsitsimärgendajate ning inimekspertide arvamused (vt ka Chlkovsky, Mihalova 2003);
- regulaarse polüseemia põhimõtete kasutamine (vt ka Buitelaar 1998);

- teatud semantiliste suhete abil grupeerimine (ko-hüponüümia, autohüponüümia jt) (Peters, Peters, Vossen 1998).

2.1.1 Semantilise peegeldamise meetod

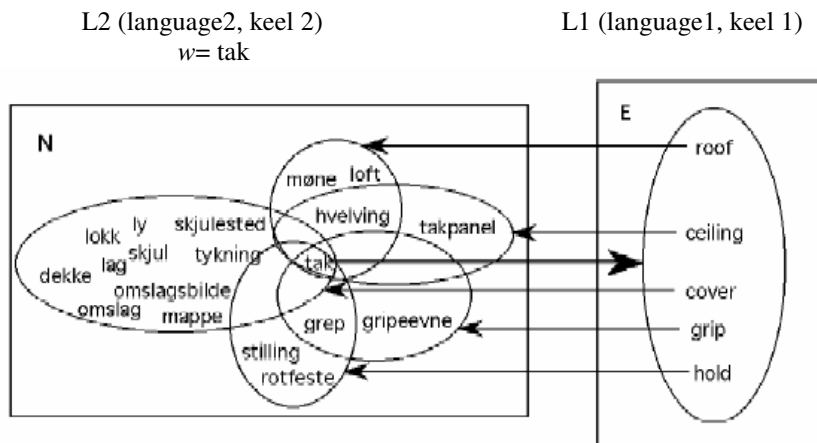
Nn semantilise peegeldamise (ingl *semantic mirrors*) meetodi on välja töötanud Bergeni ülikooli professor Helge Dyvik. Selle meetodiga on võimalik tuvastada ka semantilisi suhteid (sünonüümia, hüperonüümia, hüponüümia jt). Dyvik püüab automaatselt ekstraheerida vajalikku informatsiooni paralleelkorpusest ning väljundiks on *wordnet*-tüüpi leksikaalsemantilistes andmebaasides olevad kirjed. (Norra keelele ei ole veel loodud *wordnet*-tüüpi tesaaurust ning Dyviku eesmärk on proovida seda automaatselt luua). Siiski peab veel tulemused käsitsi üle vaatama ja hindama. (Dyvik 2002: 1–3)

Paralleelkorpusest saab ekstraheerida erinevate tõlgete kogumi iga korpuses oleva lemma jaoks. Tulemuseks on tõlkevastete keerukas võrgustik, mis ühendab kahe keele sõnavara. Seesama võrgustik lubabki kasutada ühte keelt teise peeglina kooskõlas paljude eeldustega, millest olulisimad on:

- semantiliselt lähedaselt seotud sõnadel ning sõnatähendustel on tugevalt kattuvad tõlgete kogumid;
- sõnad, millel on laiem kasutusala (või tähendus), omavad rohkem tõlkeid kui kitsa tähendusega sõnad;
- kui sõna A on sõna B hüponüüm (nt *maitsev hea* suhtes), siis tõlgetes peaks suure tõenäosusega sõna A tõlked sobima võimalike sõna B tõlgete alamhulka;
- homonüümia on ajalooline või juhuslik;
- sõnad, mis pole seotud, ei jaga omavahel tõlkeid teises keeles, välja arvatud juhul, kui jagatud tõlge on silmnähtavalt mitmene mitteseotud sõnade vahel. (Dyvik 2005: 29–32)

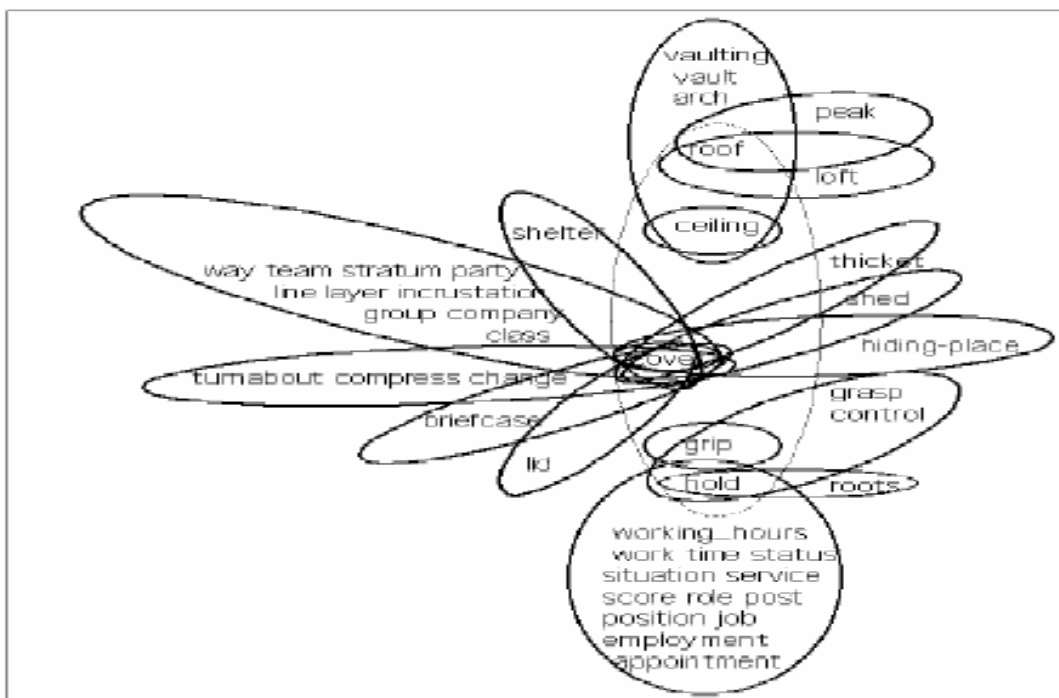
Esimene samm semantilise peegeldamise meetodi juures on homonüümide eemaldamine. Edasi luuakse nn t-kujutised (ingl. *t-image*, *translation image*). Sõna *w* tõlgete kogum hulgas L2 (keel2, *language2*) on sõna *w* esimene t-kujutis hulgas L1 (keel1,

language1). Sõna *w* esimese t-kujutise põhjal moodustatakse vastupidine t-kujutis kõikidele lõikuvatele tõlgetele hulgas L1. t-kujutisi illustreeriv joonis (Dyvik 2002: 4–6):



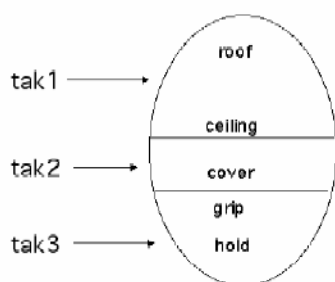
Joonis 2. Parempoolses ingliskeelses kastis on sõna *w* esimene t-kujutis, vasakpoolses norrakeelses on sõna *w* vastupidine t-kujutis.

Oluline on alles hoida kattuvad mustrid esimeses t-kujutises, et leida sõna *w* teine t-kujutis. Teine t-kujutis jaotub kolme gruppi (kitsendusena on seatud fakt, et kasutatakse vaid esimese t-kujutise liikmeid).



Joonis 3. Teine sõna w t-kujutis.

Nende kujutiste põhjal saabki jagada sõna w erinevatesse tähendustesse (Dyvik 2002: 4–6).



Joonis 4. Sõna w esimese t-kujutise tähenduste jaotus.

Järgnevalt grupeeritakse tähendused semantilistesse väljadesse (mõlemas keeles) ning tuvastatakse iga tähenduse positsioon semantilises väljas. Seda tehakse automaatselt – tunnusjoonte kogumid tuletatakse antud kogumi struktuurist. Vastavalt

komponentanalüüsile on eesmärgiks, et nn laiematel sõnatähendustel on vähem tunnusjooni ning spetsiifilisematel tähendustel on neid rohkem. Edasi saab tuletada tesauruse kirjeid, mis eeldab alammõistete, hüperonüümide, lähedaste sünonüümide ja hüponüümide kindlaksmääramist. Palju moodustub müra ja veidraid kirjeid, kuid on välja töötatud spetsiaalne meetod, kuidas vältida taolisi vajakajäämisi üleminekul *wordnet*-tüüpi andmebaasile või tesaurusele. (Dyvik 2002: 6–10)

Sellisel moel erinevate keelte tõlkeid vaadeldes on võimalik leida lisaks ka mõttekaid ühe sõna erinevaid tähendusi. Sarnase katse on läbi viinud ka Chugur, Gonzalo ja Verdejo (2002), kes kasutasid tõlkevasteid seitsmes keeles ning tulemuseks olid võimalikud sõnatähendused ning mõned semantilised suhted.

Kui eesmärgiks pidada näiteks automaatset masintõlkesüsteemi suunal eesti-inglise-eesti (nt mõnda valdkonna-spetsiifilist), siis võrreldes tõlkeid nii eesti kui inglise keeles, on võimalik moodustada just selline eesti keele erinevate sõnatähenduste komplekt, mis suurendaks oluliselt automaatse ühestaja täpsust.

2.1.2. Regulaarne polüseemia klasterdamisel

Tihti ei ole mitmetähenduslikkus juhuslik, vaid järgib teatud regulaarseid malle. Eriti laiendatakse tähendusi metonüümilisel viisil ning eelkõige puudutab regulaarne polüseemia just nimisõnu (Langemets 2004: 745). Regulaarne polüseemia (ka süstemaatiline polüseemia) ongi üks metonüümia alaliike. (Peters, Peters 2000)

Regulaarse polüseemia all mõeldakse seda, kui vähemalt kahe eri sõna tähendustel on teatav ühesugune suhe ning see regulaarsus võib avalduda nii kinnistunud tähenduste endi kui ka ühe kinnistunud tähenduse ja selle peenemate nüansside vahel (Langemets 2004: 746).

Regulaarse polüseemia definitsioon pärineb D. Апресянит: Sõna A polüseemia on antud keeles regulaarne, kui selles keeles on vähemalt üks sõna B, mille tähendusi b_1 ja

b_j eristatakse üksteisest täpselt samal viisil nagu sõna A tähendusi a_i ja a_j ning kui a_i ja b_j ega a_i ja b_j ei ole sünonüümid.

Näiteks on regulaarse polüseemia klassid LOOM-TOIT (sõna *lammas* on nii imetaja kui *liha*, mida süüakse), OBJEKT-KUJU (*püramiid*), ANUM-SISU (*pudel*), on ka kolmikuhteid jne. CoreLexi¹⁷ andmebaasis on näiteks kokku 126 sellist regulaarse polüseemia klassi.

Eraldi on käsitletud loogilist polüseemiat, mis hõlmab juhtumeid, kus sõna üldine leksikaalne kategooria ei muutu ning sõna eri tähendustel on osaliselt kattuv, sõltuv või ühine tähendus, kus nimisõna hargneb kitsamatesse alltähendustesse, võides veidi erinevates kontekstides tähistada erinevaid tähendusi. Näiteks: TOODE-TOOTJA või FÜÜSILINE OBJEKT-SELLE SISU (ta töötab *ajalehe* juures; Mari tilgutas *ajalehele* kohvi; Jüri vihastas *ajalehe* peale). (Langemets 2004: 747)

Tähendusnihke puhul on tegemist sõna üldise leksikaalse kategooria vahetumisega. Täendusnihkeid on uuritud alljärgnevate tüüpide kaupa (Langemets 2004: 747):

- 1) ASI-AINE (*tammest* parkett; *ussist* kindad)
- 2) AINE-ASI (kõvad ja pehmed *juustud*; *hirmud* ja kuidas neist jagu saada)
- 3) Pärisnimede leksikaliseerumine (otsin taskust neli *Koidulat*)
- 4) nominaalsed metonüümiad (näide juuksurisalongist: kell kümme tuleb *värv*)
- 5) fraasitähenduse nihked
- 6) uued tähendused

Regulaarse polüseemia klasse on erinevates loomuliku keele rakendustes võimalik edukalt kasutada. Antud töö jaoks on kõige olulisem, kuidas kasutada regulaarse polüseemia klasse sõnatähenduste automaatse ühestamise etapi tulemuste parandamisel. (Buitelaar 2000)

Tehakse vahet vastandatavatel sõnatähendustel (ingl *contrastive*) ning sõnatähendustel, mis on oma olemuselt täiendavad (ingl *complementary*). Teisisõnu, tehakse vahet

17 <http://www.cs.brandeis.edu/~paulb/CoreLex/overview.html> (külastatud 10.05.2007)

homonüümial ja polüseemial. Näiteks sõna *pank* on homonüümne, sõna *suu* aga polüseemne, sest omab mitut omavahel seotud tähendust. (Buitelaar 2000)

Erinevalt homonüümidest ei pea kõiki regulaarse polüseemia klassi kuuluvaid sõnatähendusi eraldi ühestama, sest sellistel sõnadel on teatud hulk tihedalt seotud tähendusi. Näiteks info-otsingu puhul võib olla piisav teada ainult seda, millisesse regulaarse polüseemia klassi sõna kuulub. Sõnatähendused ei pea olema liiga eristatud, kuid vajadusel võib ka regulaarse polüseemia reegleid järgides sõnatähendusi spetsiifilisemaks, kitsamaks muuta. Seda nimetatakse täppismääratlemata esituseks (ingl *underspecification*). Täppismääratlemata esituse all mõeldakse seda, et üksuse semantilises kirjelduses fikseeritakse üldisem, n-ö lahti harutamata tähendus, mille edasine, üksikasjalikum määratlemine lähtub kontekstist (Langemets 2005: 27). Kasutades info-otsingut näiteks sõnade *korvpall*, *jalgpall* jt võidakse otsida informatsiooni nii palli kui ka mängu kohta; sel juhul peab olema võimalus kitsendada otsingut. Samas pole ilmselt vaja eristada suhtlemise-häälitsemise vahel sõnades nagu *mõire*, *sosin* jne. (Buitelaar 2000)

Väidetakse, regulaarse polüseemia klassid on erinevates keeltes osaliselt kattuvad, on universaalseid klasse (Peters, Peters 2000). Seega võib olla kasulik uurida, kuidas mingis regulaarse polüseemia klassis olevad sõnatähendused on näiteks tõlgitud teise keelde. Nii on võimalus vaadelda, missugused sarnased sõnatähendused on võimalik grupeerida, et muuta automaatse sõnatähenduste ühestaja tööd lihtsamaks.

2.1.3. Käsitsiühestajate eriarvamuste põhjal klasterdamine

Käsitsimärgendajate eriarvamuste (vt joonis 5) põhjal on võimalik moodustada sarnaste sõnatähenduste klastreid, et vähendada erinevalt ühestatud tekstisõnade hulka (Kerner 2004: 93)

Käsitsimärgendajate eriarvamuste põhjal on võimalik vaadelda, kas tekivad mingid kindlad sõnatähenduste grupid. Kui pidevalt ei saavutata üksmeelt just mingite kindlate tähenduste osas, siis on alust oletada, et tekstikasutuses ei tule need tähendused välja, isegi kontekst ei võimalda kindlat tähendust valida. (Kerner 2004: 95)


```

piirama:925:9
tüdruku
- tüdruk+0 //_S_ com sg gen // 1 @ tüdruk:2436:3
+ tüdruk+0 //_S_ com sg gen // 3 @ tüdruk:2436:3
sisse
- sisse+0 //_D_ //
.....
.....
.....
tüdruku
- tüdruk+0 //_S_ com sg gen // 1 @ tüdruk:2436:3
+ tüdruk+0 //_S_ com sg gen // 3 @ tüdruk:2436:3
käest
  käsi+st //_S_ com sg el // 1 @ käsi:1215:5
kinni
@@ -1212,7 +1212,7 @@

```

Joonis 5. Näide erinevuste failist

Joonisel 5. on liidetud kahe käsitsiühestaja arvamused ühte faili. '+' märgiga on ühe ühestaja arvamus, '-' märgiga teise oma.

2.1.4. Semantiliste suhete abil klasterdamine

Käsitsimärgendajate erimeelsusest selgus muuhulgas, et sõnatähendusi on tekstis raskem eristada, kui sõna tähenduste vahel esineb autohüponüümia suhteid ning ko-hüponüüme või kui sünohulkade liikmed kokku langevad (Kerner 2004: 96). Kui kaks sõna tähendust jagavad sama hüperonüümi (nad on ko-hüponüümid), siis intuitiivselt on nad omavahel tugevalt seotud. Paraku ei oma selline suhe täit regulaarsust, seega nõuab selline tähenduste grupeerimine käsitsi läbivaatamist ning hindamist. (Peters, Peters 1998)

Ka autohüponüümia suhete puhul on võimalik sõnatähendusi grupeerida. Kui üks sõna tähendus on teisele hüperonüümiks, siis suure tõenäosusega moodustavad nad ühe homogeense tähenduse. (Peters, Peters 1998)

2.2. Võimalikud täiendused automaatsele sõnatähenduste ühestamise süsteemile

Järgnevalt on esitatud võimalikud lisakitsendused automaatsele ühestamissüsteemile, neid on uuritud ka töö praktilises osas.

2.2.1. Printsip „üks tähendus ühe teksti kohta“

W. Gale, K. Church ja D. Yarowsky (1992) leidsid STÜ süsteemi väljatöötamisel, et sõna tähendus sõltub ühest kindlast tekstist, täpsemalt teksti teemast. Seda nimetatakse diskursuse efektiks ja see tähendab, et kui sõna esineb kaks või enam korda ühes tekstis, siis tema tähendus on tavaliselt üks ja seesama kogu teksti piires (inglise keeles kuni 98% juhtudest). Eelnevalt mainitud autorid avastasid, et sõnatähendustel on kombeks esineda koos kindlate klastritena. Teatud sõnade puhul ei ole kaht või enam ühe sõna erinevat tähendust ühe teksti sees. Selle hüpoteesi tõestamiseks tehti erinevaid teste (nii ükskeelsete kui ka kakskeelsete ressursside põhjal) substantiividega ning jõuti järeldusele, et piisavalt suure ressursi olemasolu korral on diskursuse efekt tõepoolest tugev.

Sama printsip ilmneb ka eesti keele nimisõnade puhul ning harvaesinevate verbide puhul. Näiteks võiks tuua sõna *tüdruk*, mis esineb tekstis läbivalt ainult ühes tähenduses (ühe tähendusnumbriga). Verbidest näiteks sõnad *ärgitama*, *õngitsema*.

On kaks võimalikku rakendust, kuidas ära kasutada printsipi “üks tähendus ühe teksti kohta”:

- 1) lisakitsendusena ühestamisalgoritmides;
- 2) ei pea ühe teksti sees sõnu eraldi käsitsi ühestama, vaid saab ühe korraga antud sõnale tähendusnumbri anda.

Antud printsipi, küll pisut modifitseeritult, on rakendatud Senseval-2 võistlusel ning mõningad tulemused ja ühestamissüsteemid on ka paranenud, nende töö kiirus on tõusnud (Gale, Church, Yarowsky 1992: 234).

Fenomen ei kehti siiski adjektiivide puhul ja sagedaste verbide puhul nagu *pidama*, *saama*, *andma* jt.

2.2.2. Kontekstist tulenevad reeglid, morfoloogiline info

Ühestatud korpuse põhjal on võimalik tuvastada erinevaid mustreid, millal sõna esineb ühes või teises tähenduses, ja tuletada selle põhjal reegleid. Tihti on võimalik sõna lähikontekstis olevate teiste sõnade järgi otsustada, millises tähenduses sihtsõna on. Sageli esinevad kindlates fraasides kindla tähendusega sõnad. Samuti on märgata, et kindlad sõnavormid esinevad kindlas tähenduses, ka ainsuse ja mitmuse vormide järgi on võimalik mingile sõnale kindlat tähendust omistada (Martinez 2004: 132, 170; Marzuéz jt 2006:175). Neid reegleid on võimalik kasutada automaatses ühestamissüsteemis teatavate kitsendustena, samuti on need kasulikud käsitsimärgendajale, et valida õige või sobivaim tähendus.

Kui ka konteksti järgi ei ole võimalik tuvastada, milline sõnatähendus tekstis valida, siis on mõttekas antud sõna tähendused üle vaadata, näiteks vajadusel sarnased tähendused grupeerida (vt ka ptk 2.1.).

2.2.2.1. Grammatiseerunud vormid

Grammatiseerumise käivitaja on tähendusnihe ning selle suund on konkreetsemalt abstraktsemale (Metslang 2002: 166).

Mõnede sõnavormide substantiivikasutuse ja adverbikasutuse piiri on raske määrata, sellised on näiteks somaatilised sõnad *pea, jalg, selg, käsi*, mille vorme kolmes kohakäändes on liigitatud erinevates allikates nii substantiiviks kui adverbiks. Praegu saavad kõik sellised vormid morfoloogilises analüüsis substantiivi märgendi. Kui teatud sõnavormid kaotavad sõna täisparadigma leksikaalse tähenduse, grammatikaliseeruvad, siis tekib küsimus, et kas tuleks moodustada eraldi sõnatähendus konkreetsele vormile või on võimalik mingite formaalsete tunnuste või kontekstireeglite abil neid vajalikel juhtudel määrsõnadeks liigitada. (Vider, Muischnek 2004: 106–107)

Näiteks sõna *käsi* on lauses Pärtel võttis *kätte* ja *pesi panni puhtaks* abstraktses tähenduses, grammatiseerunud ning sellele pole enam võimalik omistada TEKsauruses esinevaid konkreetseid tähendusi. On tõdetud, et pigem on toimud tähendusnihe mingi konkreetse verbi mõjualas (Vider, Muischnek 2004: 111). Uurimist vajab ilmselt see,

milliste konkreetsete verbidega koos esinedes võib sellised grammatiseerunud vormid määrsõnadeks liigitada. Võrdluseks eelmise näitega on selles lauses sõnal *käsi* väga konkreetne tähendus: *ta valas kõigile veini ja võttis oma klaasi kätte*. Vale sõnaliigi probleem avaldub näiteks masintõlkes. Kui esimeses näitelause esitatud sõnavorm *kätte* on märgendatud substantiivina, saaks ta ilmselt ingliskeelsed tõlked kui *hand* või *arm* ning lause tõlkeks tuleks seega umbes selline: *Pärtel took into hand (arm) and washed the pan*, kui antud sõnavorm oleks märgendatud adverbina, oleks võimalik tõlge selline: *Pärtel got down on business and washed the pan*. Esimese lause puhul pole tõlge justkui päris õige, teises lauses tundub tõlge parem.

Tähenduse konkreetsuse või abstraktsuse hindamine on praegu jõukohane inimesele, empiirilise materjali detailsem uurimine annab aga ehk ka võimalusi sõnastada vajalikke reegleid arvutianalüüsi jaoks (Vider, Muischnek 2004: 111).

3. Materjali valik

Uuritavaks materjaliks on STÜ korpuse andmete põhjal moodustatud loend mitmetähenduslikest substantiividest.

Esimese etapina jaotati sõnad kahte nimekirja lähtudes sellest, milliste sõnade puhul kehtib printsiip „üks tähendus ühe teksti kohta“ ning milliste puhul mitte. Selleks eraldati polüseemsete sõnade hulgast sõnad, mis esinesid ühes tekstis rohkem kui ühes tähenduses nendest sõnadest, mis esinesid ühes tekstis ühes tähenduses. Sõnu, mille puhul kehtis printsiip „üks tähendus ühe teksti kohta“ oli kokku 436. Nende seast on eraldi ära toodud sõnad, mis esinevad rohkem kui kümnes tekstis:

- läbivalt ühes tähenduses,
- läbivalt kahes tähenduses,
- läbivalt rohkemas kui kolmes tähenduses.

Sõnu, mille puhul antud printsiip ei kehtinud, oli kokku 197.

Järgmises etapis valiti nende 197 sõna seast välja suurema esinemissagedusega substantiivid, st need sõnad, mis esinevad rohkem kui kümnes tekstis. Sõnade valikul lähtuti ka sellest, mitmes tähenduses sõna esines – sõnad, mis esinesid kahes tähenduses ainult ühe korra, jäid enamasti analüüsiosast välja. Samuti ei toodud analüüsi osasse kõiki neid sõnu, millele ei olnud võimalik juhiseid leida; kirjeldada on püütud eelkõige STÜ korpuses suure sagedusega esinevaid substantiive.

Kõik selles peatükis mainitud loendid on terviklikult esitatud lisas 3.

4. Märendusjuhiste loomine

Märendusjuhised on tuletatud sõna kasutuskonteksti uurides. Analüüsitud on 50 sagedamini esinevat substantiivi, mille kohta ei kehti printsiip „üks tähendus ühe teksti kohta“.

Iga sõna puhul otsisin kõik tema esinemised STÜ korpuses ning selle jaoks kasutasin Windowsi operatsioonisüsteemile kohandatud Unixi vahendite programmi PowerGREP¹⁸. Seejärel oli võimalik vaadelda, kas on kontekstuaalseid vms tingimusi, mille korral kasutatakse STÜ korpuses kindlaid tähendusi.

Sõnad on esitatud tähestikulises järjekorras ning iga sõna juures on esitatud:

- kõik tähendused vastavalt TEKsaurusele (paksendatud kirjas); sellel joonisel on näidatud ka
 - iga sõna tähenduse numbrit; see järgneb vahetult sõnale,
 - iga tähenduse hüperonüümi (eelmisel real enne '=>' märki),
 - olemasolul sünohulga liikmeid (esitatud kursiivis ja sulgudes),
 - vajadusel on lisatud ka sõnatähenduse seletus;
- vastavale sõnatähendusele leitud juhis(ed),
- näitelauseid STÜ korpusest (sulgudes on faili nimi, millest näitelause võetud on).

Kui on viidatud konkreetse sõnatähenduse esinemissagedusele või käsitsiühestajate eriarvamustele, siis informatsioon selle kohta on esitatud lisas 1 ja lisas 2. Kõik käsitsiühestajate eriarvamuste tabelid on võetud Kadri Kerner'i bakalaureusetööst.

¹⁸ <http://www.powergrep.com/> (15.05.2007)

4.1. Juhised

AASTA

ajavahe­mik 1, periood 1 => aasta2 ; seletus: 12 kuu pikkune ajavahe­mik, lähtepunktiks ükskõik missugune moment => aasta1 (kalendriaasta1) => aasta4 ; seletus: vanuse, ea, eluaastate kohta => aasta3 ; seletus: ajavahe­mik, mille jooksul mingi planeet (nt. Maa või Mars) teeb tiiru ümber oma päikese

JUHIS: vali aasta1:

siis, kui sõnale *aasta* eelneb vahetult ordinaalarv

...kuhu teda oli 1939. aasta lõpul tööle saadetud (tk0022).

kui sõnale *aasta* eelneb demonstratiivpronoomen *sel, sellel, tol, tollel* jt kindlale aastaarvule viitavad demonstratiivpronoomenid

...esimene laps Villem sündis tol aastal.. (tk0087)

...see oli hoopis sel aastal, kui Miina tütar linna teenijaks läks (tk0087)

JUHIS: vali aasta2:

kui sõnale *aasta* eelneb kardinaalarv

Uustulnuk pettis Kuiviku kahekümne aasta põhjal tehtud prognoose (tk0087)

Kolm aastat tagasi Keskaiglas... (tk0087)

konstruktsioonides: *aastate kaupa, pikki aastaid, palju aastaid*

..mis oli jooksnud ta peas aastate kaupa (orwell)

Kõige rohkem kahtlesid käitsiühestajad tähenduste 1 ja 2 vahel, ilmselt on tekstis raske nendel tähendustel vahet teha ka auto-hüponüümia suhte tõttu. Sõnatähenduste

analüüsil selgus, et need kaks tähendust on mõnede teatavate kontekstireeglite abil kirjeldatavad.

AEG

ajavahemik 1, periood 1
=> aeg4
=> aeg5 (<i>ajastu 2, ajajärk 2, epohh 1</i>)
näit 1, näitarv 1, lugem 1
=> aeg1 (<i>kellaaeg 1, kell 4</i>)
abstraktsioon 2, üldmõiste 1
=> aeg2
grammatiline kategooria 1
=> aeg3
mõõt 1, suurus 1, mõõde 1
=> aeg6

JUHIS: vali aeg1:

kui sõna *aeg* lähikontekstis esineb sõna *kell*

...kell näitas teist aega (tk0083)

konstruktsioonides *õigel ajal*, *õigeks ajaks*

...ja õigel ajal sööma tulla (tk0031)

Leopold jõuab vabaõhukohvikusse üsna õigeks ajaks (tk0024)

JUHIS: vali aeg2:

kui sõna *aeg* on metafoorses kasutuses kui TEGIJA

...aeg parandab kõik haavad (tk005)

...aeg astus oma astumist (tk0034)

Väga sageli on ühestatud korpuses märgendatud samades kontekstides olevaid tähendusi erinevate tähendusnumbritega; eriti palju tekib sellist märgendamist just tähenduste 2 ja 4 vahel: *kogu aeg, vähem aega, üle hulga aja, üle pika aja, tükk aega, ammu aeg ajast*. Samamoodi on sõna *aeg* saanud määrsõnafraasis *sel ajal* nii tähendusnumbri 1 kui ka 4.

Ka käsiitsiühestajate erimeelsusi arvestades on alust oletada, et selle sõna teatud tähendused on küllalt sarnased, et neid klasterdamismeetodi abil üheks tähenduseks liita.

ALGUS

ajavahe 1 => algus2 ; seletus: <i>mingi aja, perioodi v. sündmuse, tegevuse, töö jne. algusmoment v. algusosa</i> protsess1 => algus1 (kujunemine1) alustamine1 => algus3 (sissejuhatus2, alustus1, hukatus1)

JUHIS: vali algus2:

neis konstruktsioonides, milles esineb ainsuse elatiivis koos adverbiga *peale*

Suures tennisemasinas olin algusest peale liiv rataste vahel (tk0025)

Minule aga näis ta algusest peale tuttav (tk0085)

Sõnatähenduste ühestatud korpuses ei esine kordagi sõna *algus* tähenduses *kujunemine* (ehk tähendus 1), seega on ilmselt vajalikum leida reegleid, millisel juhul eristada tähendusi 2 ja 3.

ASI

olev 2
=>**asi1** (*objekt 1*)
=>**asi3**
=>**asi4** (*tehisasi 2, artefakt 2*)
tehisasi 2, artefakt 2, **asi 4**
=>**asi12** (*värk 1, tühi-tühi 1, asjad 1, asjandus 2*)
töö 3
=>**asi5** (*toimetus 2, ettevõtmine 1, ülesanne 2*)
ütlus 1, väljend 1
=>**asi6**
juht 4, juhtum 1, sündmus 1
=>**asi7** (*lugu2*)
asi iseeneses 1, idee 2, abstraktsioon 1
=>**asi8** (*nähtus2*)
tegu 2
=>**asi10**
seisund 4, olukord 4, situatsioon 1
=>**asi11**
atribuut 1, omadus 2
=>**asi9**

JUHIS: vali asi11:

ainsuse illatiivi vormis koos verbiga *puutuma*

...kuid see ei puutunud muidugi asjasse (tk0002)

Kirjanik ei puutunud asjasse (tk0130)

Ühestatud korpuses on juhtumeid, kus ühes ja samas kontekstis olevad sõna *asi* tähendused on saanud erineva tähendusnumbri: *asja selgitama* (7 ja 13), *kelle asi see on* ning *mis see sinu asi on* (5 ja 9).

Teiste tähendusnumbrite juures ei ole märgata selgeid kontekstuaalseid vms tingimusi ning arvestades ka käsitsiühendajate suhteliselt suuri erimeelsusi, on antud sõna klasterdatud peatükis 5.1.

ELU

Looduslik fenomen1

=>**elu1**; seletus: *mateeria liikumise eri vorm tema arengu teataval astmel*

olemine1

=>**elu2**; seletus: *inimese, looma ja taime füsioloogiline seisund sünnist surmani.*

=>**elu3** (*elutee1*)

olukord2

=>**elu4** (*põli1, elujärg1, elamine-olemine1, põlv3*)

abstraktsioon2

=>**elu6**; seletus: *elusorganismid kollektiivselt, täpsemalt määratlemata*

aktiivsus1

=>**elu5** (*hing3, eluvaim1, elujõud3*)

ajavahemik1

=>**elu7** (*eluaeg1*)

JUHIS: vali elu3:

kui sõnale *elu* eelneb vahetult adjektiiv *igapäevane* (tihti ainsuse illatiivis)

...kui tavaline Partei liige igapäevases elus kasutas (orwell)

...kui tõsiselt neid omadusi võeti igapäevases elus (tk0017)

kui esineb koos sõnaga *mõte*

...et omandas mõneks ajaks elu mõtte (tk0116)

JUHIS: vali elu4:

kui lähikontekstis esineb ühendverb *sisse seadma*

...olen siin oma elu sisse seadnud (post)

JUHIS: vali elu7:

koos sõnadega *terve, kogu, jooksul, kord*

Pärtel oli terve oma elu olnud puusepp ja ehitusmees (tk0020)

...et olen kogu oma elu veetnud seal.. (tk0085)

Paljud, mis mulle elu jooksul räägitakse (tk0086)

...sest olen teda kord elus näinud (tk0030)

Teatavate kontekstireeglitega oli võimalik kirjelda tähendusi 3, 4 ja 7, mis on ka kõige sagedamini esinevad sõna *elu* tähendused ühestatud sõnatähenduste korpus.

HING

elusolend1

=>**hing1** (*inimene1, inimolend1, individ1, isik1, persoon1, hingeline1*)

füsioloogiline protsess1

=>**hing2** (*hingamine1, hingetõmme1*)

aktiivsus1

=>**hing3** (*eluvaim1, elujõud3, elu5*)

miski1

=>**hing4** (*vaim3*)

=>**hing5**; seletus: (*idealistlikus filosoofias, religioonis, mütolooias*;)

mittemateriaalne surematu ja elustav alge, mis võib eksisteerida ka kehast sõltumatult

armastus1

=>**hing6** (*süda3*)

JUHIS: vali hing2:

kui sõna *hing* esineb koos verbiühenditega *kinni kõõksatama, kinni pidama, kinni jääma*

Mehel jäi hing kinni... (tk0124)

...kui Baumannid hinge kinni pidades jälgisid.. (tk0123)

JUHIS: vali hing4:

kui sõna *hing* esineb ainsuse illatiivis

Teiseks ei mahtunud talle hästi hinge, et ... (tk0111)

Hinge poeb vimm (tk00110)

kui sõna *hing* esineb ainsuse inessiivis

...nagu peituks kunstniku hinges midagi erakordselt (tk0111)

...oli Pärtli hinges iselaadi ja seletamatu rahutus (tk0020)

JUHIS: vali hing6:

kui sõna *hing* (enamasti ainsuse elatiivis) esineb sellistes konstruktsioonides kui *kogu*,
kõigest hingest

...mida me salaja kogu hingest soovisime (tk0116)

...kõigest hingest oleksin tahtnud teist aidata (tk0005)

Nii kontekstireeglite kui ka kontekstis kaasnevate sõnavormide abil oli võimalik kirjeldada sõna *hing* sagedasemini esinevaid tähendusi.

HÄÄL

füüsikaline nähtus 1, füüsiline fenomen 1

=>**hää15** (*heli1*)

=>**hää12** (*vokaliseerimine1, inimhää11*)

=>**hää13** (*häälekõla2, tämber2*)

=>**hää14** (*hää1itsus1*)

nägemus 2, seisukoht 2, arvamus 1

=>**hää11**

JUHIS: vali hää12:

kui sõna *hää1* esineb ainsuse komitatiivis

...ja vanamees hõikas kõva häälega (tk0021)

...küsis ta variserlikult kurva häälega (tk0055)

kui sõna *hää1* esineb ainsuse adessiivis

...ning küsis meelepahast väriseval häälel (tk0124)

...noomiks tasasel häälel ja sunniks andeks paluma (tk0019)

Kõige rohkem on käsitsiühendajad kahelnud tähenduste 2 ja 3 ning 2 ja 5 vahel, millest võib oletada, et nendele tähendustele oleks vaja kontekstireegleid vms, millal ühte või teist tähendust valida. Samuti muudab konkreetse tähenduse valiku raskeks sõna puhul esindatud keerukad autohüponüümia ning ko-hüponüümia suhted (vt joonis artikli alguses).

Selle sõna tähendusi on klasterdatud peatükis 5.1.

JUTT

suhtlus 2, suhtlemine 1, lävimine 1, kommunikatsioon 2

=>**jutt1** (*kõne 3, keel 5*); seletus: *verbaalne suhtlemine, see, mida keegi räägib, ütleb, jutustab*

jutustus 1

=>**jutt2** (*lugu1*); seletus: *üks eepika väikevorm*

arutelu 1, mõttevahetus 1, diskussioon 1

=>**jutt3** (*kõnelus 1, vestlus 1, jutuaajamine 1*); seletus: *inimeste omavaheline rääkimine*

vorm1

=>**jutt4** (*joon1*); seletus: *Joon, kriips, triip*

JUHIS: vali jutt1:

kui sõna *jutt* on ainsuse elatiivis

Tema jutust läks enamik kaotsi (tk0002)

...ta polnud naise jutust kõike aru saanud (tk0022)

kui sõna *jutt* laiendab kaassõna *järgi*

Juku enda jutu järgi Viljandi haiglas nii pikka kraadiklaasi polnud (tk0021)

...kus meeste jutu järgi olevat praegugi veel igasugu kraami (tk0022)

JUHIS: vali jutt2:

kui sõna *jutt* esineb koos verbiga *kirjutama*

...ning ma asusin kirjutama juttu... (tk0073)

JUHIS: vali jutt3:

kui sõna *jutt* esineb ainsuse allatiivis konstruktsioonides nagu *kellegi jutule pääsema, minema, saama*

...isiklikes asjades tahaksin mina pääseda Eduard Vilde jutule (tk0127)

kui sõna *jutt* esineb konstruktsioonis *juttu kuhugi viima*

...kuigi ma viisin uudishimutsedes kümneid kordi jutu neile radadele...(tk0073)

..viisin jutu sellele, kuidas...(tk0073)

Sõna *jutt* kõigile tähendustele on leitud teatavaid kontekstuaalseid vms reegleid, et kirjeldada konkreetset tähendust.

JÕUD

looduslik fenomen1, loodusnähtus1

=>**jõud1** (*vägi1*)

tugevus1, sitkus2

=>**jõud2** (*jaks2, ramm1, vägi2*); seletus: *kehaline tugevus*

võime1

=>**jõud3** (*jaks1, jõudlikkus2*)

põhijoon1, tunnus3, tunnusjoon1, joon3

=>**jõud4** ; seletus: mõjuv v toimiv tegur

JUHIS: vali jõud4:

mitmuse vormides

...sest tagurlikud jõud ei kavatsegi kasutada neile ettenähtud suvepuhkust (tk0055)

Mitmel juhul on üks tähendus saanud kaks võimalikku tähendusnumbrit, ka ühtlustamisprotsessis pole leitud ühte ainsat tähendust, järelikult pole olnud võimalik tekstikasutuses erinevatel tähendustel vahet teha. Näiteks lauses: *ma sulle veel haudun jõudu* on saanud sõna *jõud* märgendi 3 v 1; lauses *mind valdas kummaline nõrkus, mis imes minust jõudu* välja on sõna *jõud* saanud märgendi 2 v 3.

KEEL

kommunikatsioon 1, suhtlus 1
=> keel3
=> keel2 (<i>inimkeel1</i>)
=> keel1 (<i>kõne2, kõnekeel1</i>)
organ 2, elund 1
=> keel4
suhtlus 2, suhtlemine 1, lävimine 1, kommunikatsioon 2
=> keel5 (<i>jutt1, kõne3</i>)
rihm 1
=> keel6 (<i>piug1, piitsarihm1</i>)
asjandus1, vahend2
=> keel7 (<i>pillikeel1</i>)

JUHIS: vali keel2:

kui sõnale *keel* eelneb vahetult genitiivtribuut, mis märgib mõnd inimkeelt.

Võru keel kõdistas meeldivald kõrvu (tk0021)

..sest nii põnevaid avastusi pole soome-ugri keeltes ammu tehtud...(tk0085)

JUHIS: vali keel4:

kui sõna *keel* esineb ainsuse allatiivis

Tal tulid leelolaulud iseenesest keelele (tk0021)

...mis hingepõhjast keelele kerkib (tk0022)

kui sõna *keel* esineb sellistes konstruktsioonides nagu *keelt* või *keelega limpsama*,
limpsima, *laksutama* ning *keelt näitama*

Nüüd hakkas ülempreester keelt laksutama... (tk0089)

...kes näitab peeglile keelt (tk0086)

Käsitsiühetajad on kõige rohkem kahelnud tähenduste 2 ja 3 vahel ning kontekstireeglitega oli võimalik kirjeldada sõna *keel* tähendusnumbriga 2. Tähendus 3

esineb STÜ korpuses harva, vaid kahel korral, seega on ühestamisel võimalik toetuda statistikale ning märgendada tähendusega 3 vaid väga spetsiifilistel juhtudel. Ka oli sõnatähenduste ühestatud korpuse põhjal võimalik täheldada, et sellised konstruktsioonid, kus sõnale *keel* eelneb vahetult genitiivatribuut, esinevad väga sageli.

KEHA

looduslik objekt1

=>**keha2** (*organism1*); seletus: *inimese v. looma kogu organism*

=>**keha3** (*ilu1, liha3*); seletus: *inimese (hrv. muu elusolendi) keha (sageli hingele v. vaimule vastandatud) või selle mingi piirkond, eriti lihaosa v. välispind*

kujund1

=>**keha1**; seletus: *objekt v. kuju, millel on kõrgus, laius ja pikkus*

Selle sõna puhul ei olnud võimalik leida kontekstuaalseid vms reegleid erinevate tähenduste kirjeldamiseks. Sõnatähenduste ühestatud korpuses esineb *keha* tähendusnumbriga 2 vaid mõnel korral.

Sõna on klasterdatud peatükis 5.1. eesmärgiga vaadelda, kas tekib võimalik klaster konkreetsete tähenduste vahel.

KELL

näit1

=>**kell4** (*kellaaeg1, aeg1*)

mõõteriist1

=>**kell2** (*ajamõõtja1, ajanäitaja1*)

es1

=>**kell1**; seletus: *hrl. peekri-, pirnikujuline õõnes, tila v. vasarakesega metallese helide tekitamiseks (tihti millegi alguse v. lõpu kuulutamiseks)*

heli1

=>**kell3** (*kellahelin1, kellakõlin1*)

JUHIS: vali kell4:

kui sõna *kell* lähikontekstis esineb numeraal

Igal öösel ärkan kell kaks korduva unenäo peale... (tk0005)

Kell on pool kaksteist... (tk0087)

kui sõna *kell* esineb koos verbiga *vaatama*

...mõtles Erlend ja vaatas kella... (tk0055)

Erlend vaatas murelikult kella (tk0017)

Sõna *kell* esineb sõnatähenduste ühestatud korpuses kõige sagedamini tähendusnumbritega 2 ja 4. Sõna *kell* koos selle lähikontekstis oleva numeraaliga (ehk tähendusnumbriga 4) esineb sõnatähenduste korpuses tihti, vaid üksikutel juhtudel on tähendusnumbri 4 saanud mõnes teises kontekstis olev tähendus.

KODU

Eluase2, elamispind1

=>**kodu1**

asukoht2, koht6, asupaik2

=>**kodu2** (kodukoht1, kodupaik1); seletus: *koht, kust ollakse pärit v milles asub kellegi kodu*

JUHIS: vali kodu1:

kui sõna *kodu* on ainsuse aditiivis ning esineb koos verbidega *jõudma*, *kõndima*, *minema*, *tulema*

Kui ta koju jõudis... (tk0130)

Esimese päikeserattaga läks ta mõrul meelel koju... (tk0124)

Selle sõna puhul tuleb rõhutada, et kahes tähenduses esines ta ainult kahes tekstis ning väga harva tähendusnumbriga 2. Tihti esines sõna *kodu* just koos verbidega *jõudma*, *kõndima*, *minema*, *tulema*; seega on leitud kontekstireegel sageli rakendatav.

KOHT

koht4 (*asukoht1, paik1, asupaik1*)

=>**koht1**; seletus: *keha, eseme, hoone jne. kitsam piirkond*

(4nda taseme hüperonüüm koht4)

osa 3, ala 3, piirkond 2, regioon 2

=>**koht3** (*punkt2*)

=>**koht2**

=>**koht6** (*asukoht2, asupaik2*)

ehitus 1, ehitatu 1, ehitis 2

=>**koht5** (*pind4*)

hoone 1, ehitis 1, maja 2

=>**koht7**

positsioon 2, staatus 2

=>**koht8**

Konkreetsed kontekstireegleid antud sõna tähenduste kirjeldamiseks ei leitud.

Sellel sõnal on mitu autohüponüümia suhet, on ka ko-hüponüümia suhe; samuti on sünohulkades ning hüperonüümide hulgas palju ühesuguseid sõnu, mis teevad selle sõna tähenduste eristamise tekstis keeruliseks. Ka käsitsimärgendajate erimeelsustest selgus, et eelmainitud mitmed keerulised suhted ning ka suur tähenduste arv raskendavad märgendamist.

Antud sõna puhul on katsetatud klasterdamist peatükis 5.1., et vaadata võimalike klastrite teket.

KOOL

õppeasutus 1

=>**kool1**; seletus: *õppeasutus, kus õpilased õpetajate juhtimisel omandavad teadmisi, oskusi ja vilumusi (ka koolikollektiiv)*

näpunäide 1, õpetus 2, juhatus 2, juhtnõör 1

=>**kool2** (*õppetund1*); selgitus: *mingisuguste oskuste omandamine, koolitus; kogemused*

hoone 1, ehitis 1, maja 2

=>**kool3** (*koolimaja1, õppehoone1*); selgitus: *hoone, kus asub mingi õppe- või kasvatusasutus*

Sõna *kool* puhul ei ole selgeid kontekstuaalseid vms tingimusi, millal üht või teist tähendust valida. Käsitsimärgendajate erimeelsustest selgub samuti, et raske on valida tähenduste 1 ja 3 vahel. Nende eelmainitud tähenduste puhul on tegemist regulaarse polüseemiaga ning sõna on klasterdatud peatükis 5.2. just regulaarse polüseemia põhimõtetest lähtudes.

KORD

süsteem 3, reegliid 1
=> kord1 (<i>korraldus 7</i>)
seisund 4, olukord 4, situatsioon 1
=> kord2 (<i>korrastus 1, korraldus 5, korrapära 1</i>)
aeg 4
=> kord3 (<i>puhk2</i>)
koht 5, pind 4
=> kord4 (<i>majakorrus 1, majakord 1, korrus 1</i>)
tehisasi 2, artefakt 2, asi 4
=> kord5 (<i>kiht2</i>)
seisund 2, seisukord 1
=> kord6 (<i>korrasolek 1</i>)
viis 2, stiil 2, laad 3, maneer 1, mood 1
=> kord7
määr 1, järk 2, tase 1, aste 2
=> kord8

JUHIS: vali kord3:

kui sõna *kord* esineb partitiivis ning *-gi*-liitega

...ja talle polnud kordagi pähe tulnud, et... (orwell)

Kõige nende aastate jooksul polnud ta sellele talveõhtule kordagi mõelnud (tk0088)

kui sõna *kord* esineb ainsuse aditiivis

...ja kahel korral jällegi läbi pekstud (tk0022)

Vastasel korral poleks teda vaja (tk0111)

JUHIS: vali kord6:

kui sõna *kord* esineb koos verbiga *looma*

...ja jooksis korda looma (tk0023)

JUHIS: vali kord7:

kui sõna *kord* esineb ainsuse inessiivis ning talle eelneb vahetult inessiivis adjektiiv
(näiteks *ametlikus*, *kohustuslikus*...)

...seda filmi olime äsja kohustuslikus korras vaadanud... (tk0122)

...teen kõigile ametlikus korras teatavaks (tk0127)

Kord3 näitelause on sõna *kord* fraasis vastasel korral grammatiseerumise suunas liikuv sõnavorm. Seega oleks vajalik sõna *kord* koos talle eelneva adessiivis adjektiiviga *vastasel* liigitada hoopis adverbiks.

KÄSI

kehaliige 1, liige 1, ihuliige 1, jäse 1

=>**käsi1**; seletus: *inimese v. ahvi ülajäse randmest sõrmeotsteni*

=>**käsi5** (*käsivars 1*) seletus: *ülajäse tervikuna, õlast kuni sõrmeotsteni.*

võime 1

=>**käsi2**

pool 2, külg 2

=>**käsi3**

mängija 1

=>**käsi4**; seletus: *üks vajalikest mängupartneriist kaardimängus; bridzis ka mängija, kelle kätte mäng jääb*

JUHIS: vali käsi1:

kui esineb ainsuse aditiivis ning ei ole grammatiseerunud – nt esineb verbi *juhatama* mõjualas

Pärast tööd head kuivad kätte panna (tk0034)

...kellele raudrehasid kätte jagati (tk0083)

kui esineb ainsuse elatiivis ning ei ole grammatiseerunud vorm

...ja võtsin tal käest kinni (tk0001)

..ta pillas noa käest (tk0116)

kui esineb ainsuse inessiivis ning ei ole grammatiseerunud vorm

Ta hoiab käes lühikese varrega oda (tk0038)

...tal oli käes must ühegi kaunistuseta kleit (tk0033)

kui esineb mitmuse partitiivis

...pärast pesin mitu korda käsi (tk0030)

...lõi taas temperamentselt käsi kokku (tk0100)

JUHI: vali käsi2:

kui sõna *käsi* esineb koos verbiga *olema* sellises konstruktsioonis nagu *kätt olema*

Tema pojal oli kätt, fantaasiat ja nutti...(tk0116)

JUHI: vali käsi5:

kui sõna *käsi* esineb koos kaassõnaga *vahel, vahelt, vahele*

Ühel mihklikuisel varahommikul libises tütarlaps Tehvani käte vahelt välja... (tk0124)

Samas märkas ta, et käte vahel ei olegi enam puu jahe koor (tk0124)

...ja naine libistas end käte vahele alla (tk0083)

Käsi1 kolmes sisekohakäändes esineb palju grammatiseerunud vorme. Oluline on eristada, missuguses kontekstis on nad:

- säilitanud oma konkreetsuse, näiteks kui esineb aditiivis koos verbiga *panema*;

- grammatiseerunud, näiteks koos verbiga *paistma* (lauses: *kuni kadakatupsud tee äärest kätte paistsid*) ning peaksid saama adverbil märgendi.

Probleeme tekitavad need verbid, mille mõjualas esineb *käsi* nii grammatiseerunud kui ka konkreetsetes tähendustes, näiteks verbiga *võtma*: *ta võttis pange kätte* (konkreetne) või *nad võtsid kätte ja läksid õue* (grammatiseerunud).

LAPS

järglane 2, järeltulija 2, võsu 2
 =>**laps1**; seletus: *kellegi järglane, poeg v. tütar oma vanemate suhtes*

alaealine 1
 =>**laps2**; seletus: *inimene sündimisest sugulise küpsemise alguseni*

JUHIS: vali laps1:

kui sõnale *laps* eelneb vahetult *minu, meie* vms genitiivis personaalpronoomen

...ja tema laps elas kolm kuud meie peres (tk0005)

...elame ja töötame ju kõik oma laste heaks (tk0123)

kui lähikontekstis esineb sõna *ema, isa, vanemad, lapsevanem* (harvem ka sõnad *mees ja naine*)

...lõi isa lapse küsimuse peale käega... (tk0023)

...vanemaid laste esimese koolipäeva puhul õnnitlenud... (tk0116)

Rein Baumann oli vanemate ainus laps... (tk0123)

Laps tähendusnumbriga 1 esineb sageli STÜ korpuses just sellistes kontekstides, mis on kirjeldatavad ka analüüsil leitud reeglitega.

LAUD

mööbliese1

=>**laud1**; seletus: *rõhtsast plaadist ja ühest v. mitmest jalast vm. tugitarindist koosnev mööbliese*

=>**laud3**; seletus: *söögilaua, toidulaua kohta*

ehitusmaterjal1

=>**laud2**; seletus: *saematerjal, mille paksus on 16 -- 100 mm ja laius ületab kahekordse paksuse; sellest töödeldud profiilpuit*

JUHIS: vali laud1:

kui sõna *laud* esineb ainsuse allatiivis (mõnikord märgendatud ka kui +1, kuid selleks ilmselt vajadust pole)

Tujutult asus Leena toitu lauale seadma (tk0108)

...pani lauale suure lõhenenud külgedega leiva...(tk0108)

kui sõna *laud* esineb koos postpositatsiooniga *taha*, *taga*, *ääres*

...istun laua taha (tk0005)

Tagasi tulles peatus ta mõtlikult laua ääres (tk0023)

JUHIS: vali laud3:

kui on ainsuse aditiivis ning esineb koos verbiga *istuma* ning substantiiviga *istumine*

Seekord võidi siis ühiselt lauda istuda (tk0017)

Lapsed saadeti lauda istumiseni teise tuppa, et nad kuni lauda istumiseni jalust ära oleksid (tk00018)

Antud sõna kahte autohüponüümia suhtes olevat tähendust oli võimalik konkreetsete reeglite abil kirjeldada. *Laua* kui *ehitusmaterjali* puhul on märgatav ka võimalik tendents esineda ühes tekstis ühes tähenduses; kuid tema esinemissagedus sõnatähenduste ühestatud korpuses on veel liiga madal, et seda tähendust mainitud printsiibi alla liigitada.

LUGU

juhtum1 => lugu2 (<i>asi7</i>); seletus: <i>juhtum, sündmus v. omavahel seotud juhtumite, sündmuste kogum</i>
jutustus1 => lugu1 (<i>jutt2</i>)

JUHIS: vali lugu1:

kui sõna *lugu* lähikontekstis esineb ühendverb *valmis saama, valmis mõtlema*

...millest võib loo valmis mõelda... (tk0024)

...kui sain oma looga valmis (tk0090)

kui sõna *lugu* lähikontekstis esineb verb *avaldama* või *kirja panema*

...et mind vaevanud lugu kirja panna...(tk0112)

...ei jäta nad lugu avaldamata...(tk0122)

Sõnale *lugu* tähenduses *asi, juhtum* ei olnud võimalik leida kontekstist tulenevaid või muid reegleid.

LÕPP

joon 2, piir 4 => lõpp1 (<i>äär3</i>)
aeg2 => lõpp2 ; seletus: <i>piir, mil miski saab läbi, kauem ei kesta; lõpphetk, lõpposa</i>
piir 1, piirjoon 2 => lõpp3 (<i>piir2, äär2</i>)
lõik 2, osa 6, jaotus 2 => lõpp4 (<i>lõpuosa1, lõpposa1</i>)
juht 4, juhtum 1, sündmus 1 => lõpp5 (<i>lõpetus1</i>)
otspunkt 1 => lõpp6 (<i>ots4</i>)
seisukord 2, seisund 3, olek 1, seis 2 => lõpp7 (<i>lõppseisund1</i>)

JUHIS: vali lõpp1:

kui sõnale *lõpp* eelneb vahetult sõna *maailm* genitiivis

...sellega maailma lõppu ei jõua (tk0019)

Teiste tähenduste jaoks selgesti eristatavaid reegleid ei leitud.

Ühestatud korpus on väga sageli märgendatud sarnastes kontekstides olevaid tähendusi erineva tähendusnumbriga. Nt fraasis *lõppu tegema*, on sõna *lõpp* märgendatud nii tähendusnumbriga 5 kui ka 7.

MAA

Koht4, asukoht1, paik1, asupaik1

=>**maa1**; seletus: *selle planeedi pind kui kõige elusa asupaik vastandina atmosfäärile, taevale*

looduslik objekt1

=>**maa2** (*maismaa1, mannermaa1*)

piirkond3, territoorium1, rajoon3

=>**maa3** (*riik1*)

organisatsioon1

=>**maa4** (*riik2*)

osa3, ala3, piirkond2, regioon2

=>**maa6**; seletus: *piirkond, mis on väljaspool linnu vm asulaid*

kate1, kattekiht1, kiht3

=>**maa7** (*maapind1, maakamar1, pinnas3*)

maa8; seletus: *ruumiline vahemaa*

JUHIS: vali maa7:

kui sõna *maa* esineb ainsuse inessiivis

Maas oli rohtu ja suuri kive.. (tk0001)

...maas ei olnud mitte rohi, vaid muld (tk001)

kui sõna *maa* on ainsuse elatiivis

...võttis maast pange...(tk0001)

...kahmab maast lund...(tk0123)

kui sõna *maa* esineb ainsuse illatiivis

...tagusime need otsapidi maasse...(tk0026)

STÜ korpuses esineb *maa*⁷ kõige sagedamini, sellele on leitud ka kontekstist tulelevaid märgendusjuhiseid.

Inessiivi ja elatiivi asemel peaks semantiliselt kasutama hoopiski väliskohakäändeid, küll aga aktiveeriksid need tähenduse 6 (ala väljaspool asulaid).

MAAILM

Looduslik objekt1

=>**maailm1** (*universum1*)

olukord2

=>**maailm3**; seletus: *tingimused, milles inimene elab ja millest sõltub, kuidas inimene maailma kogeb*

elu3

=>**maailm4**; seletus: *inimeste maine elukeskkond, maapealne elu (lähtub kujutlusest, et olemas on ka surnute riik, hauatagune elu)*

inimesed1

=>**maailm2** (*üldsus1, avalikkus1*)

reaalsus1

=>**maailm5**; seletus: *maakera (kõige sellel eksisteerivaga)*

miljööl

=>**maailm6**; seletus: *keskkond, miljöö, olustik, eluring.*

JUHI: vali maailm3:

kui sõna *maailm* on ainsuse elatiivis

...vaba ja sõltumatu asjade maailmast... (tk0118)

Teiste tähenduste jaoks kontekstuaalseid vms tingimusi ei leitud. Tähendused 3 ja 5 esinevad STÜ korpuses sagedamini, seega on vajalik korpuse kasvades leida märgendusjuhendeid ka tähendusele 5.

MAJA

ehitus 1, ehitatu 1, ehitis 2

=>**maja2** (*hoone1, ehitis1*); seletus: *igasugune katuse ja seintega ehitatu*

=>**maja1** (*elumaja 1, eluhoone 1, elamu 1, eluase 1*); seletus: *ehitis, mis on mõeldud elamiseks ühele v. enamale leibkonnale*

JUHS: vali maja1:

kui sõna *maja* on ainsuse aditiivis

Karl kadus majja... (tk0005)

...kellest mõni on jäänud majja koduväiks (tk0038)

kui esineb koos verbiga *ehitama* ja noomeniga *ehitamine*

Kui oled mitmeid maju ehitanud teistele... (tk0020)

...kõigil oli majade ehitamine tähtis... (tk0020)

JUHS: vali maja2:

kui sõna *maja* on mitmuse nominatiivis

...majad seisid nüüd luitununa... (tk0020)

Vaatlusjaama majad tundusid nagu pakk-kastid.. (tk0031)

Antud sõna tähendusi oli võimalik mõningate kontekstreeglitega ning sõnavormidega kirjeldada, hoolimata tähenduste vahel olevast autohüponüümia suhtest.

MEEL

võime1	=> meel2 ; seletus: <i>vaimsete võimete kategooria; teat. liiki aistingute vastuvõtu ja eristamise võime</i> => meel4 (<i>taju1</i>); seletus: <i>meel, mille kaudu tajutakse v. tabatakse välist maailma</i>
tunne2	=> meel3 ; (<i>tuju1, meeleolu1</i>)
soov4	=> meel5 ; seletus: (<i>tunnetega ühenduses</i>) <i>tahtmine, soov</i>
loomus2	=> meel 1 ; seletus: <i>inimese loomuse, hingelaadi, vaimu, meelelaadi vms. kohta</i>
mälu1	=> meel6 ; seletus: (<i>sisekohakäänetes</i>) <i>mälu</i>

JUHIS: vali meel3:

kui esineb ainsuse adessiivis, sageli eelneb sellele vahetult ka adessiivis adjektiiv (täiendina esinev adjektiiv iseloomustab enamasti meeleolu, tuju)

...võttis meid heal meelel vastu (tk0026)

Noor mees vahtis mõrul meelel tulle (tk0124)

JUHIS: vali meel6:

kui esineb ainsuse inessiivis (välja arvatud konstruktsioonides nagu *ühes meeles*)

Unustamatult meeles on üks episood... (tk0019)

Aga selgemast selgemini on meeles, kuidas... (tk0026)

Meel6 esitatud näitelauseste põhjal võib aga oletada, et pigem on saanud inessiivis sõna *meel* öeldistäite funktsiooni konstruktsioonis *meeles olema*.

MEES

inimene 1, inimolend 1, hingeline 1, indiviid 1, isik 1, persoon 1, hing 1
=>**mees2** (*meesterahvas 1, meessoost inimene 1*)
meesterahvas 1, meessoost inimene 1, **mees 2**
=>**mees1**
=>**mees3** (*abikaasa 2*)

JUHIS: vali mees 3:

konstruktsioonides *kelle(gi) mees*; väga sageli personaalpronoomeniga.

...et ta minu meest ei tülitaks ega koormaks (tk005)

...täditütre mees töötas keevitajana (tk0130)

kui lähikontekstis esinevad verbid *abielluma, petma, lahku minema*

...kas mees teda pettis või mitte (tk0130)

...tunnete tõelist, erutavat iha mehest lahku minna (tk0127)

kui lähikontekstis esineb sõna *kodu*

Mees hakkas järjest harvemini koju jõudma (tk0005)

Mees kiskus koju jõudes sokid jalast (tk0005)

Siis sõitis lift üles ja mees tuli koju (tk0030)

JUHIS: vali mees2:

kui esineb mitmuslikes vormides, ning kui on võimalik välistada mees1 ja mees3 juhised

Kogu pikk kaldäärne jää oli mehi täis (tk0034)

JUHIS: vali mees1:

kui sõnale *mees* eelneb vahetult indefeniit-, demonstratiiv- või relatiivpronoomen

...just nimelt sellelt mehelt.. (tk0118)

...minusuguse mehega.. (tk0110)

lausetes, kus sõnad *mees* ja *naine* on rinnastatud sidesõnadega *ja, ning*

Mees ja naine kõndisid mööda ojaäärset teed (tk0083)

kui esineb lause alguses (ka osalause alguses)

Kuigi sõnal *mees* on omavahel tihedalt seotud tähendused – seda näitavad ka autohüponüümia suhted –, oli võimalik neid tähendusi kontekstireeglite ja sõnavormide kaudu kirjeldada.

MÕTE

mentaalne objekt 1, kognitiivne sisu 1

=>**mõte1**

teadmus 1, teadmine 1, tunnetus 1, kaemus 1

=>**mõte2** (*idee 1, juhtmõte 1*)

olemus 1

=>**mõte3** (*sisu 1, tähendus 2*)

tunnus 3, põhijoon 1, tunnusjoon 1, joon 3

=>**mõte4** (*tarvilikkus 1, tähtsus 3, otstarve 1*)

kõrgem tunnetusprotsess 1

=>**mõte5** (*ajutegevus 2, mõtlemine 1, mõtetegevus 1*)

JUHS: vali mõte4:

konstruktsioonides *millel, kellel*) mõtet olema (*midagi teha*)

...kas siis ongi enam erilist mõtet elada? (tk0002)

Esimesel juhul polnud mõtet hakatagi kuhugi minema... (tk0023)

Teistele tähendustele kontekstuaalseid vm tingimusi ei leitud, ilmselt on põhjus selles, et antud sõnal on abstraktseid tähendusi. STÜ korpus on ühesugustes kontekstides olevad tähendused saanud tihti erineva tähendusnumbri (nt *igas mõttes, mõtteid*

väljendama); ka käsitsimärgendajate erimeelsustest selgub, et sellel sõnal on palju sarnaseid tähendusi, mis tekstikasutuses välja ei tule.

NAINE

inimene 1, inimolend 1, hingeline 1, indiviid 1, isik 1, persoon 1, hing 1 => naine2 (<i>naisterahvas 1, naissoost inimene 1</i>) naine 2 , naisterahvas 1, naissoost inimene 1 => naine1 => naine3 (<i>abikaasa 1</i>)

JUHIS: vali naine3:

konstruktsioonides *kelle naine*; väga sageli personaalpronoomeniga.

...milline teine mees jagaks vabatahtlikult ja sõbralikult oma naist (tk0110)

...oma endist naist silmitseda (tk0024)

kui vahetult enne sõna *naine* esineb adjektiiv *endine*, *uus*

Helistab mu endise mehe uus naine (tk0005)

kui vahetult enne sõna *naine* esineb ordinaalarv

...ja soovitan tal pöörduda oma mehe esimese naise juurde... (tk0005)

...mu endise mehe kolmas naine... (tk0005)

JUHIS: vali naine1:

konstruktsioonides, kus sõnale *naine* eelneb vahetult indefiniit-, demonstratiiv- või relatiivpronoomen (teistmoodi ainult mõnel korral, ajaliselt varem ühestatud tekstides)

...oli näinud sama naist merest tulemas (tk0116)

lausetes, kus sõnad *mees* ja *naine* on rinnastatud sidesõnadega *ja*, *ning*

Mees ja naine kõndisid mööda ojaäärset teed (tk0083)

kui esineb lause alguses (ka osalause alguses)

Kuigi antud sõna puhul on tähenduste vahel autohüponüümia suhe, on siiski võimalik teatud kontekstireeglite ja sõnavormide abil tähendusi kirjeldada. Märkimisväärne on asjaolu, et sõnade *mees* ja *naine* analüüsil leitud märgendusjuhendid kattuvad peaaegu täielikult.

NÄGU

kehaosa1
=> nägu1
välimus1
=> nägu2 (<i>ilme2, pale1</i>)
avaldus2
=> nägu3 (<i>näoilme1</i>)

JUHIS: vali nägu1:

kui sõna *nägu* esineb ainsuse elatiivis

Väike punajuukseline naine õhetas näost...(orwell)

...uhkus helkis ta näost (tk0017)

Teisi kontekstireegleid vms tingimusi, millega antud sõna tähendusi kirjeldada, ei leitud. Sagedamini esinevad STÜ tähendusnumbrid 1 ja 3 ning korpuse kasvades oleks vajalik leida just nendele tähendustele rohkem võimalikke reegleid.

OLUKORD

seisund3
=> olukord2 (<i>tingimused1</i>); seletus: <i>inimeste v. asjade vm. püsiv olemiseviis v. eksistents</i>
=> olukord4 (<i>situatsioon1, seisund4</i>); seletus: <i>asjade üldine olemisviis v. mingite tingimuste kombinatsioon mingil ajahetkel</i>
situatsiooni1
=> olukord1 (<i>järg3, seisund1</i>)

JUHIS: vali olukord4:

kui esineb ainsuse elatiivis

Kentsakast olukorrast ei leia kunstnik muud pääseteed kui...(tk0122)

...ja esimene hämmeldus uuest olukorrast hakkas vaibuma...(tk0020)

kui esineb ainsuse inessiivis

Tegelikult olid naabrid ühesuguses olukorras...(post)

...kes kiljus igas võimalikus ja võimatus olukorras...(tk0031)

kui esineb ainsuse aditiivis

...et ei püüagi olukorda selgust tuua (tk0090)

...ja jäin alalõpmata piinlikku olukorda (tk0086)

Sõna *olukord* esineb sõnatähenduste ühestatud korpuses peamiselt just tähendusnumbriga 4, mille kirjeldamiseks leiti analüüsi käigus ka vastavaid sõnavorme.

OSA

Hulk3, mõõt2, määr2, kogus2, jagu1, kvantum2 => osa1 => osa2 ; (<i>tiikk1</i>)
ala4, piirkond4 => osa3 (<i>ala3, piirkond2, regioon2</i>)
objekt1, asi1 => osa4 => osa5 (<i>komponent1, koostisosa1, element3</i>)
käsitlus1, kontseptsioon1, arusaam1 => osa6 (<i>lõik2, jaotus1</i>)
osatäitmine1 => osa7 (<i>roll 1</i>)
liide1, ühendus3, seotis1 => osa8 (<i>lõik3, lüli1</i>)
kohustus1, ülesanne3, mure2 => osa9 (<i>funktisoon3, roll2</i>)

JUHIS: vali osa7:

kui sõna *osa* esineb koos verbiga *mängima*

...et inimene oli kohustatud mingit *osa* mängima...(orwell)

Ühestatud korpuses saavad samasugustes kontekstides olevad tähendused erinevad tähendusnumbrid (nt konstruktsioonis nagu *suurem osa millestki/kellestki*). Vajalikke reegleid on ilmselt raskem leida, sest tähendused on omavahel seotud (kaks autohüponüümia suhet) ning palju tähendusi on ka abstraktsed.

PABER

materjal1

=>**paber1**; seletus: *õhuke kangas- v. lehtmaterjal, mis on saadud taimse kiudaine sadestamise teel*
=>**paber4** (*paberileht1, leht2*)

dokument1

=>**paber3** (*dokument3*)

JUHIS: vali paber3:

kui sõna *paber* lähikontekstis on sõna *allkiri* või ühendverb *alla kirjutama*

Leopold kirjutab allkirja, annab paberi Annelile...(tk0024)

Olen eluaja jooksul küll mitmele paberile alla kirjutanud (tk0108)

JUHIS: vali paber4:

kui sõna *paber* esineb koos verbiga *joonistama*

Muidugi poleks ma tohtinud neid koletisi paberile joonistada... (tk0056)

Antud sõnale leiti teatavaid kontekstuaalseid tingimusi, mille puhul valida ühte või teist tähendust. Sõnatähenduste ühestatud korpuse mahu kasvades oleks vajalik leida ka selliseid reegleid, mis aitaksid eristada autohüponüümia suhtes olevaid tähendusi 1 ja 4.

PEA

kehaosa 1

=>**pea1**; seletus: *inimese v. looma keha ülemine ajude ja meeleorganitega varustatud ning kerest kaelaga eraldatud osa.*

teadmus 1, teadmine 1, tunnetus 1, kaemus 1

=>**pea2**; seletus: *pea psüühiliste protsesside ja tunnete asupaiga ning võrdkujuna*

intelligentsus 1, intelligentsus 1, arukus 1, tarkus 1, taibukus 1

=>**pea3** (*nupp 4, mõistus 2, nutt 2, aju 2*)

taimeosa 1

=>**pea4**

omadus 1

=>**pea5**

JUHIS: vali pea1:

kui esineb ainsuse partitiivis

Ta ei söandanud pead pöörata...(tk0117)

...kui Kaarel määrab juuksevedelikuga naise pead...(tk0110)

kui sõna *pea* esineb koos adverbiga *püsti*

Pea trotslikult püsti... (tk0116)

...samm oli juba üsna raske, aga pea püsti...(tk0111)

JUHIS: vali pea2:

kui sõnale *pea* järgneb vahetult adverb *läbi*

Hetkeks vilksatas Rasmusel peast läbi, et... (tk0085)

...kui mul see mõte peast läbi käis... (tk0026)

kui sõna *pea* lähikontekstis on sõna *mõte*

Ja korraga torkas mulle pähe veider mõte.. (tk0002)

...et mitte ühtegi mõtet pole peas(tkt0024)

Kõige sagedamini esinesid STÜ korpuses sõna *pea* tähendused 1 ja 2, nendele on leitud ka märgendamishendideid.

PILK

nägemine 3, vaatamine 2, märkamine 1

=>**pilk3** (*silmaavaade1, vaade4*)

=>**pilk1** (*pilguheit1*)

ilme 2, pale 1, nägu 2, nägu 3, näoilme 1

=>**pilk2**

JUHIS: vali pilk1:

kui esineb koos verbiga *saatma*

...mulle üle õla etteheitvaid pilke heites... (tk0100)

...et Liis saadab aeg-ajalt tema poole uudishimulikke pilke... (tk0024)

JUHIS: vali pilk2:

kui esineb koos verbiga *vaatama* konstruktsioonides nagu *vaatama* (*missuguse/missuguste*)
pilguga/pilkudega

...ei tohtinud vaadata süüdlase eksleva pilguga.. (tk0112)

Nad vaatasid mind veidrate pilkudega... (tk0001)

JUHIS: vali pilk3:

kui esineb koos verbiga *otsima*

Mu pilk oli otsinud laotusest punast risti... (tk0122)

...pea püsti, pilk nagu kedagi otsimas... (tk0111)

kui esineb koos verbiga *märkama*

Ta pilku märganud, astus ta tagasi vannituppa... (tk0116)

Ta oli ka ise märganud iroonilisi pilke ja muigeid.. (tk00116)

kui esineb koos verbiga *pöörama, keerama*

Ent kui ta pilgu kupli tipult ära pööras... (tk00117)

...oli see oma pilgu mujale keeranud... (tk0022)

Kuigi nii käsitsiühendajate arvates on tähendusi tekstis raske eristada (peamiselt tähendusi 1 ja 3) ning ka autohüponüümia suhe muudab konkreetse valiku komplitseerituks, on siiski võimalik teatavate kontekstireeglitega tähendusi kirjeldada.

POISS

mees2

=>**poiss1** (*poisslaps1*); seletus: *noor meessoost inimene*

laps1

=>**poiss2** (*poeg1, poja1, poju1*); seletus: (*hellitavalt:*) *pojake; väike poiss*

JUHI: vali poiss2:

kui sõnale *poiss* eelneb vahetult possessiivpronoomen *oma*

Ma tulin siia oma poisi pärast.. (tk0127)

Tarvis oma poiss ka sinna viia... (tk0127)

Kuna *laps* (täendusnumbriga 1) on hüperonüümiks *poisile* (täendusnumbriga 2), siis kehtivad *poiss* 2 puhul ka *laps* 1 reeglid.

PÄEV

ajavahemik 1, periood 1
=> päev3 (<i>tööpäev 1</i>)
=> päev5 (<i>kalendripäev 1</i>)
=> päev1 (<i>tähtpäev 1, pidupäev 1</i>)
=> päev6 (<i>valge 2, päevaaeg 1</i>)
=> päev7
aeg 4
=> päev2
ajaaühik 1
=> päev4 (<i>ööpäev 1</i>)
kiirus 1
=> päev8 (<i>päikesepaiste 2, päikene 1, päike 2</i>)

JUHIS: vali päev2:

kui sõnale *päev* eelneb vahetult aditiivivormis pronoomen *ühel*

...kui Helena ühel päeval teatas... (tk0002)

...ühel päeval märkas Erlend, et... (tk0017)

kui sõnale *päev* eelneb vahetult pronoomen *iga* (nominatiivis)

...aga nad veenavad mind iga päev... (tk0030)

...iga päev oli teda uuesti üle kuulatud... (tk0022)

JUHIS: vali päev4:

kui sõnale *päev* komitatiivis eelneb vahetult genitiivis arvsõna *üks*

...võivad end ühe päevaga pruuniks kõrvetada.. (tk0017)

...ühe päevaga ei juhtu midagi... (tk0002)

Konstruktsioonis *päevast päeva* on saanud mõlemad sõnad saanud erinevates tekstides tähendusnumbrid 2, 4 ja 5. Konstruktsioonis *järgmisel päeval* on sõna *päev* saanud erinevates tekstides tähendusnumbrid 2 ja 4.

Selgub, et kui sõnale *päev* eelneb vahetult sõna *üks*, siis vastavalt nende kahe sõna käändele on ka tähendus erinev. Siin oleks ilmselt vajalik vaadata, kas nii sarnastes kontekstides tähenduste erinev märgendamine on vajalik.

RAHU

psüühiline seisund1

=> **rahu2** (*rahulikkus2, meelerahu1*); seletus: ärevuse v. stressi puudumine

suhe1

=> **rahu3**; seletus: riiu, tüli, vaenu puudumine, leplikkus

situatsioon1

=> **rahu1** (*rahuaeg1*); seletus: riikide vahekord, kus kumbki riik ei rakenda teise suhtes vägivalda

kõrgendik1

=> **rahu4** (*kari2*); seletus: veealune v. pisut üle veepinna ulatuv murdlainetusega kivine kõrgendik meres

JUHS: vali rahu1:

kui sõna *rahu* lähikontekstis esineb sõna *sõda*, *taplus* vms

...sõda ei ole rahu... (orwell)

JUHS: vali rahu2:

kui sõna *rahu* esineb ainsuse komitatiivis

Kõuts jätkas kõigutamatu rahuga oma teekonda... (tk0085)

...näeme suurejoonelist liikumist ühe-aegselt küpsemise vaikse rahuga... (tk0038)

Oli võimalik leida teatud reegleid, et kirjeldada konkreetseid tähendusi. *Rahu* kui *kõrgendik* on homonüüm, selle tähendus erineb teisest kolmest oluliselt.

RIIE

riideese2
=>**riie1** (*rõivas1, riidees1, riietusesel, rõivaese1*)

asi4, tehisi2, artefakt2
=>**riie2** (*kangas1, tekstiil1*)

JUHIS: vali riie1:

kui sõna *riie* esineb mitmuse vormides

Riideid polnud ta seljast ära võtnud... (tk0030)

Need hirmupuhtad riided panid mind väristama... (tk0056)

JUHIS: vali riie2:

kui sõna *riie* esineb ainsuse elatiivis

...kallist riidest ja hea rätsepa õmmeldud... (tk0085)

...maavillasest riidest vateeritud jopp... (tk0022)

Mõlemale tähendusele on leitud üks sõnavormist tulenev reegel, millisel juhul valida konkreetset tähendust. *Riie2* materjalisõnana ei moodustagi mitmuse vorme, käsitsiühestajale on selle tähenduse eritamine selge, küll aga saab antud juhust rakendada automaatne sõnatähenduste ühestaja.

RUUM

abstraktsioon2
=>**ruum3**; seletus: *piiritu kolmemõõtmeline avarus, milles kõik asub*
=>**ruum1**; seletus: *liikumisruum*

pind4
=>**ruum4**; seletus: *hoone, ehitise, sõiduki vm. sisemus v. selle sisemuse vaheseintega eraldatud osa*
=>**ruum2** (*tuba1*)

JUHIS: vali ruum1:

kui sõna *ruum* esineb koos verbidega *tegema*, *andma*

...et anda ruumi teistele... (orwell)

Lutrinile tehti ruumi... (tk0017)

Antud sõna tähenduste vahel on kaks autohüponüümia suhet, mis muudab nende kirjeldamise komplitseerituks. Leiti ainult üks kontekstist tulenev reegel, STÜ kasvades oleks vaja reegleid ka teistele tähendustele, kuivõrd tekstis esinevad kõik tähendused enam-vähem võrdse sagedusega.

RÕÕM

emotsioon1

=>**rõõm1** (*rõõmutunne1*)

nauding1

=>**rõõm2** (*lust2, lõbu1*)

JUHIS: vali rõõm1:

kui sõna *rõõm* esineb koos sõnaga *tundma*

...ja sellegi üle tundis jaak rõõmu (tk0108)

...ta tundis hingematvat rõõmu... (tk0116)

Selle sõna tähendused on abstraktsed ning tähendused on suhteliselt sarnased ning seega ei olnud võimalik leida palju kontekstireegleid.

SAMM

üksikliigutus1

=>**samm1** (*ast1*)

heli1 1

=>**samm2**; seletus: *jala v. jalgade siirmisliigutusel kostev heli käimisel, jooksmisel, tantsimisel vm. liikumisel.*

taktika1

=>**samm3** (*ettevõtmine3, teguviis1*)

JUHIS: vali samm1:

kui sõnale *samm* eelneb vahetult pronoomen *mõni*

Mõne sammu koos edasi läinud... (tk0124)

Leena tegi mõned sammud, et võõrast saatma minna... (tk0108)

kui sõna *samm* on ainsuse adessiivis ja sellele eelneb sobiva tähendusega adjektiiv

Mees ja naine kõndisid aeglasel sammul... (tk0083)

...veidi väsinul sammul... (tk0085)

TAEVAS

Väljamõeldud koht1

=>**taevas1** (*paradiis1*)

looduslik objekt1

=>**taevas2** (*taevalaotus1, laotus1*)

JUHIS: vali taevas1:

kui sõna *taevas* esineb võrdlusena koos sõnaga *kingitus*

Aga siis ilmus kui taeva kingitus naabrite kuuene poisike (tk0021)

kui sõna *taevas* esineb konstruktsioonis nagu *tänu taevale*

Tervis on praegu parem, tänu taevale (tk0021)

Taevas tähenduses *paradiis* esineb sõnatähenduste ühestatud korpuses harva, küll oli võimalik seda teatud kontekstireeglitega kirjeldada.

TEE

joon 2, piir 4
=>**tee1**(*siht2, orientatsioon2, kurss1, suund2*)
=>**tee6**(*liikumistee1*)
=>**tee7**
jook1
=>**tee5**
tegu2
=>**tee2**(*viis 1, vahend 1, abinõu 1, kanal 1, meede 1, mood 3*)
tehisasi 2, artefakt 2, asi 4
=>**tee4**(*rada2, teerada1*); seletus:*miski, mida mööda saab minna ühest kohast teise*
=>**tee3**; seletus: *avalikuks kasutamiseks (reisimiseks v. vedudeks) avatud tee*

JUHIS: vali tee1:

kui sõna *tee* esineb koos verbidega *jätkama, näitama*

...siis jätkas loom oma teed (tk0085)

...ja kuuhelek näitas teed (tk0108)

JUHIS: vali tee4:

kui esineb koos liikumisverbidega (*sõitma, vonklema, minema* jt)

...sinna tuli sõita pikk tee elektrirongi ja trammiga (tk0025)

...tee vonkles ümber igerike põõsaste (tk0031)

Sõnal *tee* on nii autohüponüümia suhteid kui ko-hüponüüme, mis muudab sellistes suhetes olevad tähendused tekstis raskemini eristatavaks. Käsitsiühestajad kahtlesid kõige rohkem tähenduste 3 ja 4 vahel ning viimast oli võimalik mitme lähikontekstis oleva sõna abil ka kirjeldada.

TUND

ajaiühik1 => tund1 ; seletus: 1/24-le päevast vastav ajaiühik
ajavahemik1 => tund2 (koolitund1)

JUHIS: vali tund2:

kui sõna *tund* lähikontekstis (vähemalt sama lause piires) esineb sõna *klass*, *õpetaja*, *kool* (ka liitsõnaosisena)

...kui kirjanduseõpetaja tutvustas ühes tunnis kirjanike pärandi säilitamist... (tk0130)

Ei ole võimalik algklassides tunde anda... (tk0017)

Sõna *tund* esineb sõnatähenduste ühestatud korpuses tähendusnumbriga 2 just siis, kui selle lähikontekstis esineb mõni kooliga seotud sõna; ehk leitud reegel katab ära suurema osa *tund* 2 esinemised.

TUNNE

psüühiline nähtus 1 => tunne2 (tundmus2) => tunne3 (tundeelamus 1, emotsioon 1, tundmus 1)
kõrgem tunnetusprotsess 1 => tunne1 (tunnetamine 1, taju 3)

Antud sõna puhul ei olnud võimalik leida kontekstuaalseid ega muid tingimusi, millal üks või teine tähendus esineb. Sõnad on tekstikasutuses raskesti eristatavad (ka käsitsiühestajad ei erista tihti kõiki tähendusi) ning ka intuiitiivselt väga sarnased. Seda sõna on klasterdatud peatükis 5.1.

TÖÖ

tegevus 1, tegutsemine 1, toimetus 1, toiming 2

=>töö3

=>töö5 (*toimine1, töötamine1, funktsioneerimine1*)

=>töö2 (*ülesanne1, operatsioon1, toiming1*)

kohustus 1, ülesanne 3, mure 2

=>töö1 (*tööülesanne1, töökohustus1*)

produktid 1, saadused 1, tooted 1, toodang 2

=>töö4 (*töötulemus1*)

tegevusala 1, tegevusvaldkond 1, kutse 2, ala 1

=>töö6 (*ametikoht1, amet2, töökoht3*)

JUHIS: vali töö6:

tihti siis, kui sõna *töö* esineb ainsuse allatiivis

Järgmisel päeval tööle minnes veendus ta... (tk0017)

...kuhu teda oli 1939.aasta lõpul tööle saadetud (tk0022)

tihti siis, kui sõna *töö* esineb ainsuse ablatiivis

...kui töölt koju tulime... (tk0019)

tihti siis, kui esineb ainsuse adessiivis

...et proua Mannke on alles teisipäeval tööl (tk0100)

Kõige sagedamini esinevad STÜ korpuses tähendused 1 ja 3 ning ka käsitsiühestajad on olnud kõige suuremal eriarvamusel just nende kahe tähenduse puhul. Samuti on kaheldud tähenduste 1 ja 6 vahel, ning viimasele tähendusele on leitud ka sõnavormidest tulenevaid reegleid.

UKS

tõke 1, takistus 1, tõkestus 1

=>**uks1**; seletus: *hingedel v. rullidel liikuv takistus, mis sulgeb pääsu mingisse ruumi*

auk 3, ava 2, avaus 1

=>**uks2** (*ukseava2*); seletus: *avaus seinas, mille kaudu saab mingisse ruumi siseneda ja sealt väljuda; avaus, mida suletakse uksega (1. täh)*

JUHIS: vali uks1:

kui sõnale *uks* järgneb kaassõna *taga, tagant*

...et ta seisab selle ukse taga... (tkf0123)

...aga miks pidi tema oma magamistoas ukse taga tammuma (tkf0123)

kui sõna *uks* esineb koos verbiga *avama, avanema*

Ukse avanud, tundis ta... (tkf0123)

Vannitoa uks avanes... (tkf0116)

kui esineb koos verbiga *koputama*

Pärast koputas keegi uksele... (tkf0123)

Paar tundi tagasi oli ta direktori uksele koputanud... (tkf0112)

kui sõna *uks* esineb koos adjektiiviga *suletud*

On veel vara, uks on suletud (tkf0123)

JUHIS: vali uks2:

kui sõna *uks* esineb ainsuse adessiivis (tavaliselt koos verbiga *seisma*)

Leaanika seisis endiselt uksele... (tkf0033)

...kui automaatidega fritsud uksele seisis... (tkf0022)

Sellised konstruktsioonid nagu *uksest sisse* ja/või *välja* liikuma on saanud nii tähendusnumbri 1 kui ka 2. Antud sõna on klasterdatud peatükis 5.2., sest sõna kuulub regulaarse polüseemia klassi.

UNI

Füsioloogiline seisund1
=>**uni1**; seletus: *puhkamise olek*

kujutlemine1
=>**uni2** (*ulm1, unenägu1*)

JUHIS: vali uni1:

kui esineb ainsuse partitiivis koos verbiga *tulema* (tihti eitavas kõneviisis)

Ilmselt ei tulnud Habemikul und (tk0022)

...aga meil ei tulegi und (tk0110)

JUHIS: vali uni2:

kui esineb koos verbiga *nägema*

Pulmaööl nägi Rein unes, et... (tk0123)

Nägin unes lõvisid (tk0030)

STÜ korpuses esinevad mõlemad tähendused enamvähem võrdselt ning leitud reeglitega on võimalik kirjeldada ka teatud tähenduse esinemist.

VARI

kaitsekiht1
=>**vari1** (*ulualune1*)

fenomen1
=>**vari2**; seletus: *valguse takistuse tõttu tekkinud tumedam kogu*

JUHIS: vali vari1:

kui sõna *vari* esineb ainsuse aditiivis

Lastekodusse ta ei varju joosta ei saanud... (tk0019)

Leena puged varsti müüri varju süngi... (tk0108)

Antud sõna esineb STÜ korpuses veel suhteliselt harva, seega ei olnud võimalik leida rohkem reegleid.

öö

Ajavahemik1

=>**öö1** (ööaeg1, pime1); seletus: *päikeseloojaku ja -tõusu vaheline ajavahemik*

=>**öö2**; seletus: *magamiseks mõeldud aeg*

JUHIS: vali öö1:

kui lähikontekstis esineb sõna *päev* (tihti just rinnastavalt)

Nõnda möödus mitu päeva ja ööd (tk0124)

Jah, öö ja päev pole tõesti vennad (tk0030)

kui sõnale *öö* eelneb vahetult adverb *kogu, terve*

Hingede virvatuled jäid kogu ööks vilkuma (tk0021)

Pärast istus kogu öö kase all (tk0124)

kui esineb konstruktsioonis nagu *öö (ööd) on (olid) missugune(sed)*

Öö oli sedakorda jahedam (tk0124)

Öö oli soe (tk0118)

Hoolimata kahe tähenduse vahelisest autohüponüümia suhtest oli võimalik leida kontekstist tulenevaid reegleid, et kirjeldada tähendust.

5. Klasterdamine

5.1. „Semantilise peegeldamise“ meetodi kasutamine klasterdamisel

„Semantilise peegeldamise“ teel on klasterdatud valitud sõnu, mille kohta eelmises analüüsi osas ei leitud või leiti vähe kontekstuaalseid vms tingimusi, mille korral üks või teine tähendus esineb. Lähtutakse asjaolust, et kui pole võimalik sõnatähendusi teatud reeglite abil kirjeldada, siis võivad need tähendused olla liiga sarnased, et neid tekstis eristada.

Analüüs on pigem eksperimentaalne, püütakse näidata, et teatud sõnade puhul on võimalik ning mõttekas tekstides sarnaseid sõnatähendusi mitte eristada.

Lisaks tõlkevastete kasutamisele on mõnel juhul arvesse võetud ka käsitsimärgendaja eriarvamustest saadud infot ning vajadusel tuuakse klasterdamisse ka auto-hüponüümia suhted ning ko-hüponüümia.

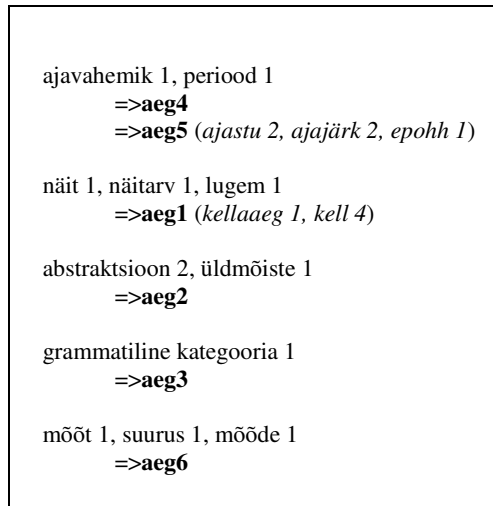
„Semantilise peegeldamise“ meetod on antud analüüsis kohendatud TEKsauruses olevate ingliskeelsetele sünohulkade vastetele, paralleelkorpusi pole kasutatud. Samuti pole püütud tuvastada erinevaid suhteid nagu nt hüperonüümia, hüponüümia jne.

Sõna juures on esitatud:

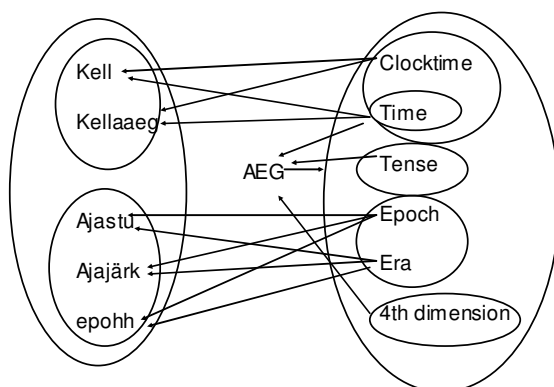
- tema kõik tähendused vastavalt TEKsaurusele ning sellel joonisel on näha ka võimalikud autohüponüümia suhted ja ko-hüponüümid;
- vastavalt vajadusele ka käsitsimärgendajate eriarvamused; esimese veerus on sõna ning need tähendused, milles on käsitsiühendajad eriarvamusel olnud ning teises veerus on esitatud antud eriarvamuste esinemise sagedus;
- „semantilise peegeldamise“ joonis; nooltega on näidatud tõlkevasted ning ringides on võimalikud sarnased tähendused (sünohulgad) (joonisel pole näidatud kõikide sõnade võimalikke, vaid ainult vastava sõna jaoks olulisi tõlkeid);
- „semantilise peegeldamise“ tulemusel saadud uus tähendusjaotus, sulgudes on ka näidatud, millised tähendused moodustasid klatri;

- illustreerivad testtõlked inglise keelde, et vaadata, millisel juhul on klasteri moodustamine olnud mõttekas; laused on võetud sõnatähenduste korpusest ning sõna taga on STÜ korpuses olev vastav tähenduse number.

AEG klasterdamine



Sõnal *aeg* on vastavalt TEKsaurusele 6 erinevat võimalikku tõlget ning järgmisel joonisel on näidatud, milline tähendus millise(d) tõlke(d) saab.



Vastavalt „semantilise peegeldamise“ meetodit rakendades tekib viis võimalikku tähendust:

- (1,2,4)AEG (kui: ajavahemik, üldmõiste, kellaaeg) (inlg *time, clocktime*)
- (3)AEG (verbi grammatiline kategooria) (ingl *tense*)
- (5)AEG, ajastu, ajajärk, epohh (ingl *era, epoch*)
- (6)AEG (mõõde, koordinaat) (ingl *4th dimension*)

Ka käsitsiühendajad ei ole tihti eristanud just tähendusi 2 ja 4 ning 1 ja 4, mis moodustasid „semantilise peegeldamise“ tulemusel sarnaste sõnatähenduste klatri.

<i>sõna;tähendus</i>	<i>sagedus</i>
aeg;2,4	21
aeg;1,4	8
aeg;4,6	8
aeg;4,5	6

Illustreerivad testtõlked

Naine võis veeta aega (2) kohvikus: woman could spend time, clocktime in the caffee

...kui aeg (4) on käes: ...when the time, clocktime comes.

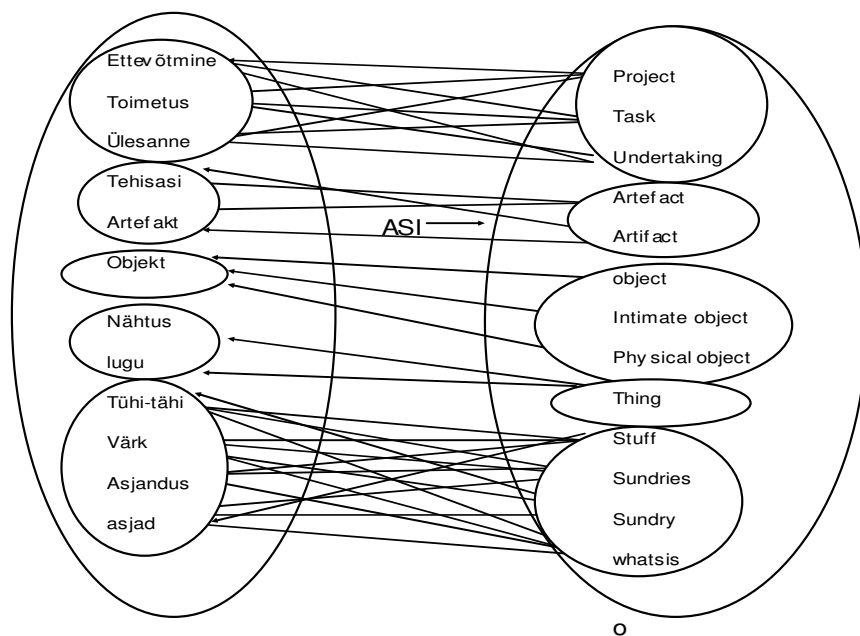
...õigel ajal (1) sööma tulla: to come to eat at the right time, clocktime

Ingliskeelsele vastele *clocktime* on võimalik anda selles sünohulgas tõlkena näiteks väiksem tähtsus või kaal; eelistatum tõlge oleks ingliskeelse sõnaga *time*.

ASI klasterdamine

olev 2	=> asi1 (<i>objekt 1</i>)
	=> asi3
	=> asi4 (<i>tehisasi 2, artefakt 2</i>)
tehisasi 2, artefakt 2, asi 4	=> asi12 (<i>värk 1, tühi-tähi 1, asjad 1, asjandus 2</i>)
töö 3	=> asi5 (<i>toimetus 2, ettevõtmine 1, ülesanne 2</i>)
ütlus 1, väljend 1	=> asi6
juht 4, juhtum 1, sündmus 1	=> asi7 (<i>lugu2</i>)
asi iseeneses 1, idee 2, abstraktsioon 1	=> asi8 (<i>nähtus2</i>)
tegu 2	=> asi10
seisund 4, olukord 4, situatsioon 1	=> asi11
atribuut 1, omadus 2	=> asi9

Vastavalt TEKsaurusele on sõna *asi* sisaldavatel sünohulkadel 14 võimalikku tõlget inglise keelde, need on näidatud järgmisel joonisel:



„Semantilise peegeldamise“ tulemusena moodustub viis võimalikku tähendust sõnale *asi*.

- (5) ASI, ettevõtmine, ülesanne, toimetus (ingl *project, task, undertaking*)
 (4) ASI, tehisasi, artefakt (ingl *artefact, artifact*)
 (3,6,7,8,9,10,11) ASI, lugu, nähtus (ingl *thing*)
 (1) ASI, objekt (ingl *intimate object, object, physical object*)
 (12) ASI, tühi-tähi, värk, asjandus, asjad (ingl *stuff, sundries, sundry, whatchamacallit, whatsis*)

Käsitsiühetaajate erimeelsustest on näha, et kõige rohkem kaheldakse tähenduste 7 ja 11 vahel ning ka 3 ja 11, 9 vahel.

<i>sõna+täh</i>	<i>sagedus</i>
asi;7,11	10
asi;3,11	7
asi;3,9	7

Illustreerivad testtõlked

Ega auto pole elus kõige tähtsam asi (3): the car isn't the most important thing in life

...sest ta kirjutas päris igasuguseid asju (6): because he wrote all kinds of things

Tema jaoks oli see suur asi (7): for him it was a great thing

Kirjandus on väga tore asi (8): literature is a great thing

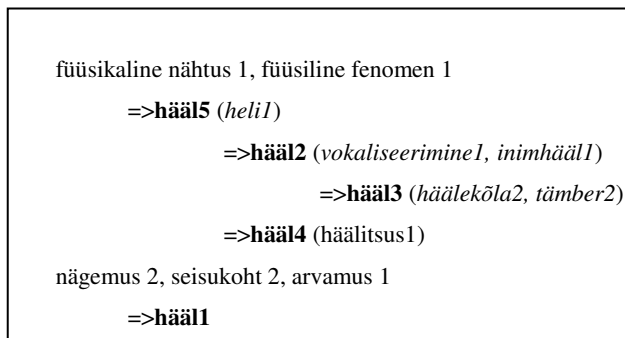
Teise inimese armastus oli veel ainus asi (9): the love of another person was the only thing

Kuidas ta küll võis niisuguse asjaga (10) hakkama saada?: how could he do such thing?

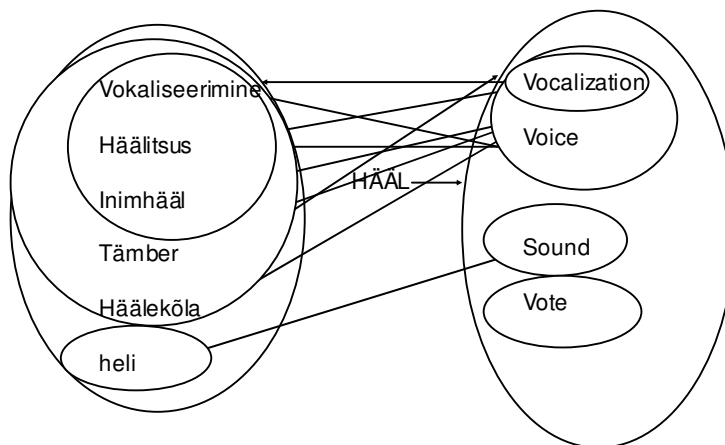
Tegelikult oli asi lausa vastupidi (11): actually the thing was vice versa

Kõikidele nende tähendustele sobib tõlkevastena *thing*; ka käsitsiühetaajate eriarvamustest selgub, et neid tähendusi kasutatakse samades kontekstides, seega võrreldes inglise keelega oleks klatri moodustamine õigustatud.

HÄÄL klasterdamine



Vastavalt TEKsaurusele on sõnal *hää1* neli võimalikku tõlget, mis on järgmised:



„Semantilise peegeldamise“ tulemusena tekib neli võimalikku tähendust:

- (2,4) HÄÄL, vokaliseerimine, hää1itsus, inimhää1 (ingl *voice, vocalization*)
- (3) HÄÄL, tämber, hää1lekõla (ingl *voice*)
- (5) HÄÄL (kui heli) (ingl *sound*)
- (1) HÄÄL (kui arvamus, hää1etamine) (ingl *voice*)

Klastri moodustasid tähendused 2 (kui inimehää) ja 4 (kui häälitus), mis algselt olid ka ko-hüponüümia suhtes. Nende tähendustega on tugevalt seotud ka tähendus 3 (kui tämber, häälekõla); võimalik, et eelmainitud klastris olevad tähendused on sellele hüperonüümiks.

Käsitsiühendajate erimeelsustest selgus, et kõige raskemini eristatavad olid tähendused:

<i>sõna+täh</i>	<i>sagedus</i>
hää;2,3	6
hää;2,5	4

Siin ei ühti täielikult peegelmeetodi tulemusel saadud klaster ning käsitsiühendajate erimeelsustest tulenev võimalik klaster. Tõlgetest selgub, et moodustatud klaster tähenduste 2 ja 4 vahel on õigustatud, kuivõrd mõlemad tähendused saavad sobiva ingliskeelse vaste.

Illustreerivad testtõlked

Ei tohi eriti kõva häälega (2) laulda: You can't sing with a loud voice; you can't sing with loud vocalization

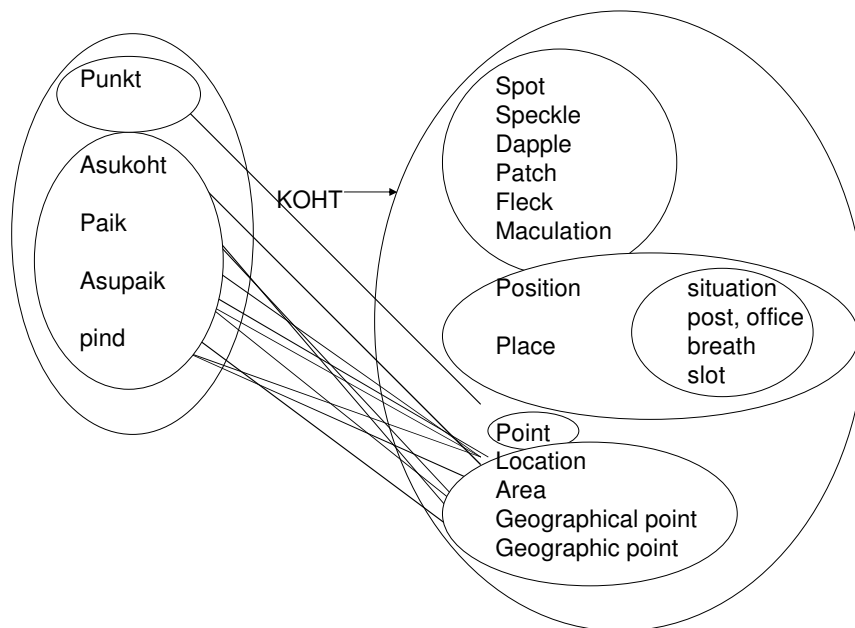
Tema hää (3) oli käskiv: his voice was demanding; his vocalization was demanding

Metsast kostis linnu kurblik hää (4): the voice of a bird which came from the woods was sad; the vocalization of a bird which came from the woods was sad

KOHT klasterdamine

koht1; seletus: *keha, eseme, hoone jne. kitsam piirkond*
 osa 3, ala 3, piirkond 2, regioon 2
 =>**koht3** (*punkt2*)
 =>**koht2**
 =>**koht6** (*asukoht2, asupaik2*)
koht4 (*asukoht1, paik1, asupaik1*)
 ehitus 1, ehitatu 1, ehitis 2
 =>**koht5** (*pind4*)
 hoone 1, ehitis 1, maja 2
 =>**koht7**
 positsioon 2, staatus 2
 =>**koht8**

Vastavalt TEKsaurusele on 18 võimalikku tõlkevastet:



„Semantilise peegeldamise“ tulemusena tekib neli võimalikku tähendust:

- (1,7) KOHT (ingl *spot, speckle, dapple, patch, fleck, maculation*)
- (2,8) KOHT (ingl *position, place, situation, post, breth, slot, office*)
- (3) KOHT, punkt (ingl *point*)
- (4,5,6) KOHT, asupaik, pind, paik, asukoht (ingl *location, area, geographic(al) point*)

Käsitsiühestajate eriarvamusest on näha, et kaheldakse paljude tähenduste vahel:

<i>sõna+täh</i>	<i>sagedus</i>
koht;3,4	3
koht;1,4	2
koht;2,+1	2
koht;3,+1	2
koht;5,+1	2
koht;5,6	2
koht;1,7	1
koht;1,8	1
koht;2,6	1
koht;2,7	1
koht;2,8	1
koht;3,6	1
koht;4,5	1
koht;4,8	1
koht;6,8	1
koht;8,+1	1

Illustreerivad testtõlked

Ta leidis käepideme juures ühe terava koha (1): he found a sharp spot near the handle

Mul ei tule võõras kohas (7) und: i can't get sleep in a strange spot

Mees istus juhi kohal (2): the man was sitting on the drivers place, position, situation, slot

Tuli võttis Killi koha (8) kodus ära: fire took away Kill's place, position, situation, slot at home

Nad seisisid väga käidava koha (4) peal: the stood on a very public location, area, geographic point

Köögi koht (5) on piisavalt suur; the kitchen area, location, geographic point is pretty large

Koht (6) hakkas meile järjest enam meeldima: the area, location, geographic point started to impress us even more

Testtõlgetest selgub, et klatri võiksid moodustada tähendused 2 ja 8, teiste klatriite puhul on ilmselt vajalik vaadata tõlkeid just paralleelkorpusdest ning kindlasti on oluline leida sobivad autohüponüümiasuhted ja võimalikud ko-hüponüümid.

KEHA klasterdamine

looduslik objekt1

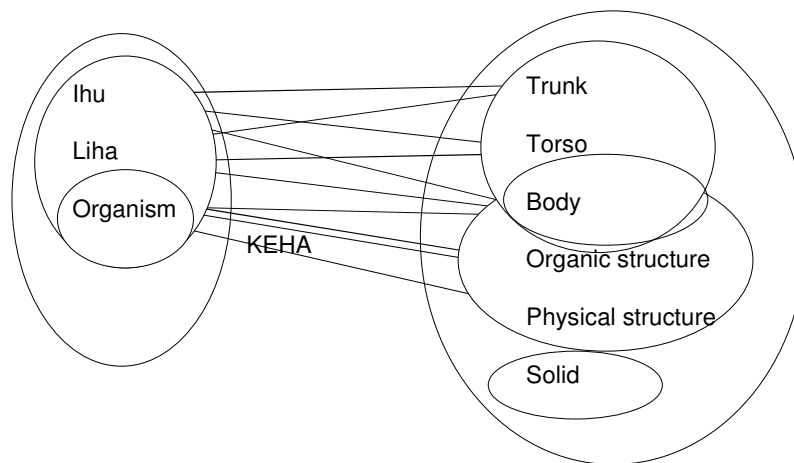
=>**keha2** (*organism1*); seletus: *inimese v. looma kogu organism*

=>**keha3** (*ilu1, liha3*); seletus: *inimese (hrv. muu elusolendi) keha (sageli hingele v. vaimule vastandatud) või selle mingi piirkond, eriti lihaosa v. välispind*

kujund1

=>**keha1**; seletus: *objekt v. kuju, millel on kõrgus, laius ja pikkus*

Sõnal on vastavalt TEKsaurusele kuus võimalikku tõlget:



„Semantilise peegeldamise“ tulemusena tekivad võimalikud tähendused:

(1) KEHA (ingl *solid*)

(2,3) KEHA, ihu, liha, organism (ingl *body, trunk, torso, organic structure, physical structure*)

Illustreerivad testtõlked

Nende kehad (2) topiti särkidesse: their bodies, trunks, torsos, organic structures, physical structures were put into shirts

Lõhnastab hoolikalt oma võrratut keha (3): she carefully perfumes her magnificent body, trunk, torso, organic/physical structure.

Testtõlgetest selgub, et tähendusi 2 ja 3 on kõige parem tõlkida ilmselt inglise keelde sõnaga *body*, seega on võimalik antud klaster ka moodustada. Ka viitab võimalikule klastrile nende kahe tähenduse-vaheline auohüponüümia suhe. Harvem võib tekkida olukord, kus eristamist vajab *keha* kui *rinnak*, ilma käteta, kaelata jms.

TUNNE klasterdamine

psüühiline nähtus 1

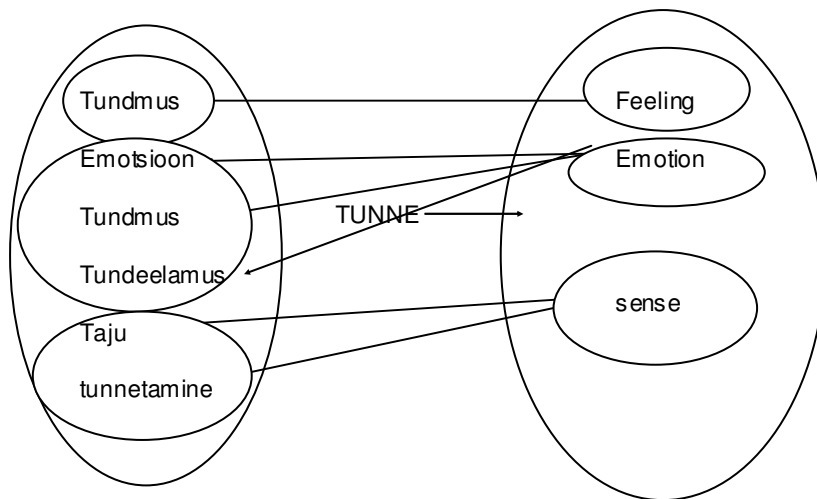
=>**tunne2** (*tundmus2*)

=>**tunne3** (*tundeelamus 1, emotsioon 1, tundmus 1*)

kõrgem tunnetusprotsess 1

=>**tunne1** (*tunnetamine 1, taju 3*)

Sellel sõnal on TEKsauruse järgi kolm võimalikku tõlget:



Nagu näha ka eelmiselt jooniselt, ei ole võimalik antud andmestiku põhjal klastreid moodustada, kõik tähendused säilitavad võrdluses inglise keelega oma tähtsuse. Küll aga on käitsiühestajate erimeelsustest näha, et tekstikasutuses on siiski tähendused raskesti eristatavad ning võiksid moodustada teatavad klastrid:

<i>sõna+täh</i>	<i>sagedus</i>
tunne;2,3	10
tunne;1,2	6
tunne;1,3	1

Illustreerivad testtõlked:

Mul pole kunagi olnud nii kurb tunne (1) lahkuda: I've never had such a sad sense, feeling, emotion to leave

Tundsin meeldivat tunnet (2): I felt a pleasant feeling, emotion, sense

Ma ei saa sellest tundest (3) lahti: I can't get rid of this emotion, feeling, sense

Testtõlgetest selgub, et ingliskeelsed vasted *emotion* ja *feeling* sobivad kõigile tähendustele, vaste *sense* moodustab omaette spetsiifilisema tähenduse. Antud juhul on siiski võimalik moodustada klaster tähenduste 2 ja 3 vahel. Võimaliku klatri moodustumist näitab ka nende kahe tähenduse vaheline autohüponüümia suhe.

5.2. Regulaarse polüseemia põhimõtete kasutamine klasterdamisel

Regulaarse polüseemia põhimõtteid kasutades on klasterdatud kolme valitud sõna. Eesmärgiks on näidata, et teatud regulaarsetesse polüseemia klassidesse kuuluvaid sõnatähendusi on võimalik klasterdada.

Oluline on vahet teha tähendustel, mis moodustavad kindla klatri ning tähendustel, mida on vajadusel võimalik ka kitsamalt kirjeldada (täppismääratlemata tähendustel). Sõnatähenduste klasterdamisel on kasutatud ingliskeelseid vasteid, et vaadata võimalikku tähenduse sobivust klastrina.

Analüüsil nimetatud regulaarse polüseemia klassid on võetud Wim Petersi doktoritööst (2004) ning kohandatud vastavalt eesti keelele. Edaspidi on võimalik eesti keele nimisõnade regulaarse polüseemia käsitlust laiendada Margit Langemetsa valmiva doktoritöö¹⁹ põhjal.

Iga sõna juures on esitatud:

- tema kõik tähendused vastavalt TEKsaurusele;
- regulaarse polüseemia klass;
- vajadusel käsitiühestajate eriarvamused.

¹⁹ Doktoritöö teema "Nimisõnade regulaarne polüseemia eesti keeles"

KOOL

õppeasutus 1
=> kool1 ; seletus: <i>õppeasutus, kus õpilased õpetajate juhtimisel omandavad teadmisi, oskusi ja vilumusi (ka koolikollektiiv)</i>
näpunäide 1, õpetus 2, juhatus 2, juhtnõör 1
=> kool2 (<i>õppetund1</i>); selgitus: <i>mingisuguste oskuste omandamine, koolitus; kogemused</i>
hoone 1, ehitis 1, maja 2
=> kool3 (<i>koolimaja1, õppehoone1</i>); selgitus: <i>hoone, kus asub mingi õppe- või kasvatusasutus</i>

Kuulub regulaarse polüseemia klassi: HOONE-ÕPPEASUTUS (organisatsioon, institutsioon).

Läbi ILI-lingi saab *kool* kui õppeasutus ingliskeelse tõlke *school*; *kool* kui hoone saab võimalikeks tõlgeteks nii *school* kui ka *schoolhouse*. Kuna ka käsitsimärgendajad ei suuda tekstis sageli neid kahte tähendust eristada ning tegelikus tekstikasutuses ei pruugi need ka välja tulla, siis oleks mõttekas need kaks tähendust grupeerida.

<i>sõna+täh</i>	<i>sagedus</i>
kool;1,3	8
kool;1,2	2

Samasugusesse klassi kuuluvad ka näiteks sõnad *haigla, ülikool, ministeerium, kontor*.

KÜLG

osa1
=> külg2 (<i>pool2</i>); seletus: <i>eseme üks osa</i>
pind2
=> külg1 (<i>tahk2</i>); seletus: <i>eseme v maastikuobjekti üks (väline) osa</i>
kehaosa1
=> külg3 ; seletus: <i>inimese v. looma vasak- v. parempoolne kehaosa, eriti roiete piirkond</i>

Kuulub regulaarse polüseemia klassi FÜÜSILINE OBJEKT-ASPEKT.

TEKsauruses puudub sõna *külg* aspekti tähenduses, seega vaatasin antud tähenduse võimaliku tõlke inglise keelde keeleveebis²⁰ olevatest inglise-eesti sõnastikest ning läbiv võimalik tõlge oleks *side*. Sama sõnaga tõlgitakse inimese *külge*. Võrreldes inglise keelega on võimalik need kaks tähendust klasterdada, kuid siiski peaks jääma võimalus teatud juhtudel neid tähendusi kitsamalt eristada, näiteks ükskeelses info-otsingus.

Samasugusesse regulaarse polüseemia klassi kuulub nt sõna *tahk*.

UKS

tõke 1, takistus 1, tõkestus 1

=>**uks1**; seletus: *hingedel v. rullidel liikuv takistus, mis sulgeb pääsu mingisse ruumi*

auk 3, ava 2, avaus 1

=>**uks2** (*ukseava2*); seletus: *avaus seinas, mille kaudu saab mingisse ruumi siseneda ja sealt väljuda;*

avaus, mida suletakse uksega (1. täh)

Kuulub regulaarse polüseemia klassi TÕKE-AVAUS

Mõlemad tähendused saavad vastavalt TEKsaurusele ingliskeelse tõlkevaste *door*; *uks* kui *avaus* saab lisaks võimalikke tõlkeid (*doorway*, *room access* jt). Need viimatimainitud tõlked on ilmselt marginaalsemad, seostuvad pigem ainult sünohulgas oleva sõnaga *ukseava*. Ka käsitsiühendajad ei erista väga tihti neid kahte tähendust ning ilmselt ei tulegi need tekstikasutuses iga kord välja ning seega on võimalik need tähendused klasterdada. Analoogselt võib toimida sõnaga *aken*.

²⁰ keeleveeb.edu.ee (10.05.2007)

6. Tulemused, järeldused

Selles peatükis antakse ülevaade eesti sõnatähenduste korpuse analüüsil tekkinud võimalikest praktikas rakendatavatest tulemustest ning tehakse järeldusi analüüsitud materjali põhjal.

Sõnad, mille puhul kehtis printsiip „üks tähendus ühe teksti kohta“

Analüüsi esimeses etapis vaadeldi, milliste nimisõnade puhul kehtib printsiip „üks tähendus ühe teksti kohta“ ning milliste puhul see ei kehti. Tulemusena tekkis nimekiri sõnadest, mille puhul printsiip ei kehti ning nimekiri sõnadest, mille puhul see kehtib. Nende sõnade puhul, mille tähendused järgisid seda printsiipi, oli samuti võimalik fikseerida, millistes konkreetsetes tähendustes need sõnad esinesid. Selgus, et oli sõnu, mis esinesid läbivalt (s.t. mitmes erinevas tekstis) ainult ühes tähenduses (nt *õlg*, *sõber*). Eraldi nimekirjadena on ära toodud sõnad, mis esinesid läbivalt kahes tähenduses (nt *mood*, *tüdruk*) ning sõnad, mis esinesid läbivalt rohkem kui kolmes tähenduses (vt Lisa 3). Sõnatähenduste ühestatud korpuse kasvades on võimalik jälgida, kas antud printsiipi järgivate sõnade nimekirja lisandub uusi liikmeid või kas mõni sõna sellest nimekirjast hoopis kaob.

Printsiip „üks tähendus ühe teksti kohta“ kehtis mõnede homonüümsete nimisõnade hulgas (nt *õlg*, *põlv*, *nutt*); küll jäid sellest nimekirjast välja sõnad nagu *reis*, *viis*, *tee*. Selle põhjuseks on ilmselt asjaolu, et viimati mainitud sõnadel on lisaks homonüümidele ka veel suhteliselt palju polüseemseid tähendusi (näiteks on sõnal *tee* TEKsauruse järgi kokku 7 tähendust ning paljud neist on polüseemsed). Analüüsi tulemusena võib nentida, et printsiipi „üks tähendus ühe teksti kohta“ ei järgi sõnad, mille polüseemiatase on kõrge, selles nimekirjas on üksikuid sõnu, millel on neli tähendust (nt *suund*, *kasv*). Tavalisimad sõnad, mille puhul antud printsiip kehtib, omavad kahte tähendust (nt *tahe*, *häbi*) või harvem ka kolme (nt *isik*, *toit*). Printsiip ei kehti enamasti ka nende sõnade puhul, mille esinemissagedus tekstides on suurem. Seega leiab kinnitust tõsiasi, et sõnad, millel on palju

tähendusi ning mille esinemissagedus on ka suurem, pole automaatsel sõnatähenduste analüüsil hästi kirjeldatavad selliste heuristiliste meetoditega, nagu „üks tähendus ühe teksti kohta“.

Printsiipi „üks tähendus ühe teksti kohta“ järgivate sõnade rakendused

Printsiipi „üks tähendus ühe teksti kohta“ on kasutatud teadmistel põhinevates automaatsetes sõnatähenduste ühestamissüsteemides ning lähtuvalt sellest printsiibist peab automaatne süsteem ühestama vastava sõnatähenduse tekstis ainult ühe korra, edasi lisab ta samadele sõnadele automaatselt juba ühestatud tähenduse. Automaatsele ühestajale on vajalik omada ka informatsiooni selle kohta, millised printsiipi järgivad sõnad esinevad nt läbivalt ainult ühes tähenduses (või ka rohkemates tähendustes), sest see hõlbustaks süsteemi tööd veelgi. Tulemuslikuks võib osutuda printsiibi „üks tähendus ühe teksti kohta“ kasutamine just valdkonna-spetsiifilisi tekste ühestades, näiteks ehitusalastes tekstides saaks sõna *klaas* pidevalt just materjali tähenduse, mitte joogiklaasi tähenduse.

Ka käsitsimärgendajad saavad antud printsiibi järgimisest kasu. Käsitsimärgendaja juhendis on see printsiip juba ära märgitud, kuid nüüd lisandub sinna ka nimekiri konkreetsetest sõnadest²¹.

Kontekstist ning teatud sõnavormidest tuletatud märgendamisjuhised

Neid nimisõnu, mille puhul printsiip „üks tähendus ühe teksti kohta“ ei kehti, on automaatsel süsteemil seega raskem ühestada. Sellised sõnad on tavaliselt kõrge polüseemiatasemega ja tekstides suure esinemissagedusega, mis tegelikult muudab oluliseks just nende sõnatähenduste kirjeldamise – millisel juhul on sõna mingis konkreetses tähenduses.

²¹ Käsitsiühestamise täiustatud juhend asub internetis, <http://www.zone.ee/yhestamisjuhend/index.htm> (30.05.2007)

Ühe võimaliku lahendusena oli vaadeldud selliste sõnade lähikonteksti STÜ korpuses, ehk leida selliseid lähikontekstis olevaid sõnu, mis mõjutavad sihtsõna tähendust. Näiteks, kui sõnade *mees* või *naine* lähikontekstis esines nimisõna *abielu*, olid nii *mees* kui *naine* saanud abieluinimese tähenduse; ja näiteks kui sõna *elu* lähikontekstis esines ühendverb *sisse seadma*, oli sõna *elu* saanud alati *elujärje*, *elamise-olemise* tähenduse. Mõnede sõnade puhul oli täheldatav, et sellele vahetult eelnev või järgnev sõna mõjutas uuritavat tähendust, näiteks kui genitiivis sõnale *jutt* järgnes vahetult kaassõna *järgi*. Ka kindlas konstruktsioonis esinemine mõjutas sageli tähendust, nt *kätt olema* või vaatama *missuguse/missuguste pilguga/pilkudega*. Paaril juhul on võimalik märgata, et lause või osalause alguses asetsemine annab sõnale mingi teatud tähenduse (nt *mees*, *naine*). Mõne sõna konkreetset tähendust mõjutas ka näiteks konkreetne kääne, milles sõna esines, näiteks sai sõna *hing* ainsuse allatiivi käändes alati ühe kindla tähenduse. Mõne sõna puhul toodi välja ka grammatiseerunud tähendusi ning võimalusel näidati, missugused kontekstuaalsed vms tingimused aitavad otsustada, kas tähendus on grammatiseerunud või mitte. Näiteks sõna *käsi* konstruktsioonis *kätte juhatama* on muutunud abstraktseks ning seda peaks märgendama hoopis adverbina.

Osa juhustest on rakendatavad ka erinevate sõnade erinevate tähenduste puhul – kui üks tähendus on teise suhtes hüperonüüm, siis on võimalik selle juhised kasutada ka hüponüümi puhul. Juhise sõnastuses peab vahetama vaid sõnad. Näiteks sõna *laps* (tähenduses *järglane*) on hüperonüümiks sõnale *poiss* (tähenduses *järglane*, *kellegi laps*, *poeg*).

Kõige vähem kontekstuaalseid jms juhiseid oli võimalik leida nende sõnatähenduste kohta, mis on oma loomult abstraktsed (nt *asi*, *meel*, *mõte*) ning tihti ka metafoorses kasutuses (nt *aeg*). Abstraktsed nimisõnad on need, mida ei saa katsuda, näha, kuulda ning mis väljendavad tundeid, mõtteid, omadusi. Kuna abstraktseid tähendusi on raske tekstis ühestada ning tihti kuuluvad nad ka sagedamini esinevate mitmetähenduslike nimisõnade hulka, on ehk mõttekas edaspidi neid veelgi detailsemalt uurida. Näiteks LDOCEs²² on

²² <http://www.longman.com/ldoce/> (10.05.2007)

Clark²³ järgi abstraktseid sõnu rohkem kui konkreetseid. Üks võimalusi oleks kindlaks teha kõige sagedasemini esinev tähendus (Gale jt 1992), eriti efektiivne võiks see olla nende sõnatähdenduste ühestamisel, mille puhul on mõni tähendus tööpoolest märgatavalt sagedasem kui teine. Abstraktsetele nimisõnadele on võimalik anda teatud keeletehnoloogilistes rakendustes väiksem kaal; näiteks info-otsingul rõhutakse päringus olevale konkreetsemale nimisõnatähdendusele rohkem ehk konkreetsem tähendus on otsingu puhul olulisem ja tähtsam.

Selliseid kontekstireegleid või mõne sõnavormi esinemist ühes kindlas tähenduses saab ühestamisel kasutada lisakitsendusena automaatne sõnatähdenduste ühestamissüsteem. Juhendatud masinõppega ühestamissüsteemid kasutavad konteksti- vms reegleid; sellist automaatset süsteemi treenitakse ühestatud näidete põhjal ning mingit teksti ühestades saab ta omandatud informatsiooni edukalt kasutada.

Ka käsitsiühestajad saavad kasutada analüüsi tulemusena loodud märgendamisjuhiseid, et hõlbustada erinevate tähenduste vahel valimist. Magistritöö ühe tulemusena loodud märgendusjuhised on lisatud käsitsiühestamise juhendisse.

Sarnaste sõnatähdenduste klasterdamine

Valitud sõnadega, mille tähendusi ei olnud kontekstireeglitega vms võimalik kirjeldada, katsetati käsitsi sarnaste sõnatähdenduste klasterdamist. Kasutati nn semantilise peegeldamise meetodit, käsitsimärgendajate eriarvamustest saadavat informatsiooni, erinevate suhete abil klasterdamist ning regulaarse polüseemia põhimõtteid. Erinevaid meetodeid kombineeriti, et saada veelgi paremaid ja täpsemaid tulemusi.

Selgus, et kõrge polüseemiatasemega sõnadel on tihti sarnaseid tähendusi, mida tekstikasutuses ei ole võimalik eristada, ning mida teatavate keeletehnoloogiliste rakenduste jaoks pole isegi mõtet eristada. Näiteks sõna *aeg*, mille puhul käsitsimärgendajad ei suuda valida tekstis tähenduste *ajavahemik* ja *üldmõiste* vahel ning ka semantiline peegeldamine inglise keelega ei tuvasta nende sõnade vahel erinevust. Sõnatähdenduste ühestamist

²³ <http://www.csrc.uqam.ca/en/SiteGraham/stats.html#concrete-abstract> (10.05.2007)

keerulisemaks muutvaid autohüponüümia ja ko-hüponüümia suhteid on võimalik teatud juhtudel klasterdada – kui tähendused on tekstikasutuses täiesti sarnased, näiteks sõna *hää*l puhul, kus *hää*l tähenduses *inimhää*l on hüperonüümiks *hää*lele tähenduses *täm*ber, on ilmselt põhjendatud nende kahe tähenduse klasterdamine.

Kui klasterdamismeetodeid kasutades ilmneb sarnaseid tähendusi, mille puhul on võimalik neist klaster moodustada, siis võib analüüsil loodud märgendusjuhistega kirjeldada ka suuremat hulka tähendusi.

Kindlasti oleks „semantilise peegeldamise“ meetodi rakendamine paralleelkorpusete ning reaalse teksti põhjal kasulik, kuivõrd oluline on just sõnetähenduste kasutus loomuliku keele tekstides.

Kui on teada, et mingid kindlad sõnad kuuluvad kindlatesse regulaarse polüseemia klassidesse, siis on võimalik teatud juhtudel sellised sõnad klasterdada üheks tähenduseks, näiteks sõna *kool* nii *asutuse* kui *ehitise* tähenduses; vajadusel võib jääda võimalus näidata tähendusi ka kitsamalt, eraldi. Näiteks sõna *kül*g, mis on nii füüsiline objekt kui aspekt, küll aga võib teatud rakenduses olla oluline just mingi eseme (*välise*) *kül*je eristamine *aspekti* tähendusest. Regulaarse polüseemia täpsema uurimise abil on võimalik leida nn täppismääratlemata sõnad, mille puhul oleks võimalik vajadusel tähendust kitsamalt kirjeldada (nt *jalgpall* on kas *pall* või *mäng*). Kasulik oleks teada ka neid regulaarse polüseemia klasse, mille puhul sellist kitsamat lisatähendust pole vajalik eristada (nt *kool* kui *ehitis* ja *institutsioon*). Kui alusleksikonis on kõik sõnadevahelised võimalikud regulaarsed suhted kirjeldatud, on võimalik neid teatud mustrite järgi näiteks klasterdada või vajadusel kasutada täppismääratlemata esitusi.

Mida vähem on alusressursis tähendusi, mille vahel valida, seda kiiremaks muutub ka ühestamistöö. Treumuth ja Kikas (2007) kirjutavad, et kõiki sõnu ühestades võtab lause analüüs keskmiselt 2 minutit aega. Veelkord on võimalik näidata valdkonna-spetsiifilist sõnatähenduste ühestamist kui üht võimalikku lahendust – kui alusressurssi on koondatud ainult mingi teatud valdkonna sõnavara ning olulised semantilised suhted, siis on automaatsel sõnatähenduste ühestajal vähem tähendusi analüüsida ning töökiirus peaks olema oluliselt suurem.

Sarnaste sõnatähenduste klasterdamine peaks toimuma automaatselt, sest käsitsi muutub see ülesanne ajamahukaks, kuid saadud uusi tähendusjaotusi peab ilmselt siiski üle vaatama inimekspert.

Kokkuvõtvalt tulemustest, mis on rakendatavad automaatse ühestamissüsteemi töös:

- sõnade loend, mille puhul kehtib printsiip „üks tähendus ühe teksti kohta“; kitsamalt ka, missugustes tähendustes sellised sõnad esinevad;
- analüüsitud mitmetähenduslike nimisõnade formaliseeritud märgendusjuhised, et hõlbustada konkreetse sõnatähenduse valikut.

Kokkuvõtvalt tulemustest, mida saab rakendada käsitsi sõnatähenduste ühestamisel:

- sõnade loend, mille puhul kehtib printsiip „üks tähendus ühe teksti kohta“; kitsamalt ka, missugustes tähendustest sellised sõnad esinevad;
- analüüsitud mitmetähenduslike nimisõnade formaliseeritud märgendusjuhised, et hõlbustada konkreetse sõnatähenduse valikut.

7. Kokkuvõte

Käesolevas magistritöös pakuti erinevaid võimalusi automaatse sõnatähenduste ühestamise täiustamiseks.

Töö käigus analüüsiti mitmetähenduslikke nimisõnu. Lähtekohaks võeti tõsiasi, et neid sõnu, mille puhul ei kehti printsiip „üks tähendus ühe teksti kohta“ on raskem automaatselt ühestada. Moodustati kaks nimekirja sõnadest – need, mille puhul antud printsiip kehtib (436 sõna) ning need, mille puhul ei kehti (197).

Järgmises uurimistöös etapis analüüsiti sõnatähendusi, mille puhul eelmainitud printsiip ei kehtinud. Selliseid sõnatähendusi kirjeldati kontekstireeglite abil ning ka morfoloogiliste tunnuste abil. Analüüsiti kokku 50 sõna, lähemalt kirjeldati sõnade valiku kriteeriume peatükis 3. Kokku leiti analüüsitud sõnadele 122 märgendusjuhist, mis abistaksid nii automaatset sõnatähenduste ühestajat kui ka käsitsi sõnatähenduste ühestajaid.

Kolmas analüüsi osa oli pigem eksperimentaalne; näidati, et erinevate meetodite abil on võimalik sarnaseid sõnatähendusi klasterdada, et muuta automaatse sõnatähenduste ühestaja tööd lihtsamaks – tõsta selle täpsust ning töökiirust. Käsitsi klasterdati valitud sõnu, mille tähendusi ei olnud võimalik kontekstireeglitega vms tingimustega kirjeldada või millel oli sellised reegleid vähe.

Analüüsi käigus selgusid ka konkreetsed ja rakenduskõlbulikud tulemused nii automaatse sõnatähenduste ühestamise süsteemi kui ka käsitsi sõnatähenduste ühestaja jaoks, neist on pikemalt juttu eelmises peatükis.

Võimalikud edasised uurimissuunad. Töö käigus tuvastati, et automaatne sõnatähenduste ühestamine saavutaks täpsemaid tulemusi, kui alusressurss oleks loodud kindla valdkonna jaoks. Samuti on vaja alusressursi granuleeritust kohandada vastavalt kindlale keeletehnoloogilisele rakendusele. Selleks oleks vaja muuhulgas leida kõige parem

meetod, või kombineerida erinevaid meetodeid, et klasterdada sarnaseid sõnatähendusi just vastava valdkonna või keeletehnoloogilise rakenduse eripärasid silmas pidades.

Kirjandus

- Agirre, Eneko; Edmonds, Philip. Introduction 2006. – Word Sense Disambiguation. Ed E. Agirre, P. Edmonds. Netherlands: Springer, Series: Text, Speech and Language Technology, Vol. 33, pp 1–22.
- Budanitsky, Alex; Hirst, Graeme 2001. Semantic Distance in WordNet: An experimental, application oriented evaluation of five measures. Proceedings of the NAACL Workshop on WordNet and Other Lexical Resources, Pittsburgh, USA, pp 24–34.
- Buitelaar, Paul 1998. CoreLex: An Ontology of Systematic Polysemous Classes. In: Formal Ontology in Information Systems. IOS Press, Amsterdam, pp 14–19.
- Brockmann, Carsten; Lapata, Mirella 2003. Evaluating and combining approaches to selectional preference acquisition. Proceedings of the European Association for Computer Linguistics (EACL), Budapest, Hungary, pp 27–34.
- Buitelaar, Paul 2000. Reducing Lexical Semantic Complexity with Systematic Polysemous Classes and Underspecification. In: Proceedings of the ANLP2000 Workshop on Syntactic and Semantic Complexity in Natural Language Processing Systems, Seattle, USA; <http://www.dfki.de/~paulb/anlp00.html> (15.05.2007).
- Chklovski, Tim; Mihalcea, Rada 2003. Exploiting Agreement and Disagreement of Human Annotators for Word Sense Disambiguation. – Proceedings of the Conference on Recent Advances in Natural Language Processing. Borovetz, Bulgaria, pp 4–12; <http://www.isi.edu/~timc/papers/ranlp2003.pdf> (16.05.2007).
- Chugur, Irina; Gonzalo, Julio; Verdejo, Feliza 2002. Polysemy and sense proximity in the senseval-2 test suite. In Proceedings of the ACL-2002 Workshop on "Word Sense Disambiguation: recent successes and future directions", pp 32–39.
- Dyvik, Helge 2002. Translations as Semantic Mirrors: From Parallel Corpus to Wordnet. In 23rd International Conference on English Language Research on Computerized

- Corpora of Modern and Medieval English (ICAME), Gothenburg, Sweden, pp 1–13; <http://www.hf.uib.no/i/LiLi/SLF/Dyvik/ICAMEpaper.pdf> (12.05.2007).
- Dyvik, Helge 2005. Translations as a semantic knowledge source. In *Proceedings of The Second Baltic Conference on Human Language Technologies*, pp 27–38.
- Gale, William; Church, Kenneth; Yarowsky, David 1992. One Sense Per Discourse. – DARPA Workshop on Speech and Natural Language, New York, pp. 233–237.
- Hwee Tou Ng, Hian Beng Lee 1997. DSO Corpus of Sense-Tagged English Linguistic Data Consortium, Philadelphia.
- Ide, Nancy; Véronis, Jean 1998. Introduction to the special issue on word sense disambiguation: the state of the art. – *Computational Linguistics*, No 24, pp. 2–40.
- Ide, Nancy; Erjavec, Tomaz; Tufis, Dan 2002. Sense Discrimination with Parallel Corpora. In *Proceedings of the ACL SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*, pp 54–60.
- Kaljurand, Kaarel 2004. Sõnatähenduste automaatne määramine eestikeelsetes tekstides. Arvutiteaduste Instituudi seminar. Tartu, käsikiri; <http://math.ut.ee/~kaarel/NLP/Publications/ATI2004/wsd.pdf> (17.05.2007)
- Kerner, Kadri 2004. Sõnatähendused tekstides ja tesauruses ühestajate erimeelsuste põhjal. Käsikiri. Bakalaureusetöö.
- Kilgarriff, Adam 1997. “I don’t believe in word senses”. – *Computers and the Humanities*, No 32 (2), pp. 91–113.
- Leacock, Claudia; Chodorow, Martin; Miller, George 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24 (1), pp 147–165.
- Lesk, Michael 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone, ACM SIGDOC '86, The Fifth International Conference on Systems Documentation, Proceedings of ACM Press.
- Langemets, Margit 2004. Mõnda nimisõnade semantikast. - *Keel ja Kirjandus* nr.10

- Langemets, Margit 2005. Generatiivse leksikoni teooria ja generatiivsuse piirid. Tartu Ülikool, Keeleteaduse ja –tehnoloogia doktorikool, Teoreetiline Keeleteadus Eesti 2, teesid, lk 26–28.
- Marquéz, Lluís; Escudero, Gerard; Martínez, David; Rigau, German 2006. Supervised Corpus-Based Methods for WSD. – Word Sense Disambiguation. Ed E. Agirre, P. Edmonds. Netherlands: Springer, Series: Text, Speech and Language Technology, Vol. 33, pp 167–216.
- Martínez, David 2004. Supervised Word Sense Disambiguation: Facing Current Challenges. PhD Thesis. Universidad del País Vasco.
- McCarthy, Diana; Koeling, Rob; Weeds, Julie; Carroll, John 2004. Finding predominant senses in untagged text. In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain.
- McCarthy, Diana 2006. Relating WordNet senses for word sense disambiguation. In Proceedings of the ACL Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together, Trento, Italy pp 17–24; <ftp://ftp.informatics.susx.ac.uk/pub/users/dianam/relateWN.pdf> (15.05.2007).
- Metslang, Helle 2002. Grammatikalisatsiooniteooria eesti keele tasul. – Teoreetiline keeletadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 4. Toim R. Pajusalu, I. Tragel, T. Hennoste, H. Õim. Tartu, lk 164–179.
- Mihalcea, Rada, Chklovski, Timothy. Teaching Computers 2004. Building Multilingual Linguistic Resources with Volunteer Contributions over the Web, in The LISA Newsletter - Globalization Insider.
- Müürisep, Kaili, 2000. Eesti keele arvutigrammatika: süntaks. Dissertationes Mathematicae Universitatis Tartuensis 22. Tartu.
- Navigli, Roberto 2006. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance. Proceedings of COLING-ACL 2006, Sydney, Australia pp 105–112; http://www.dsi.uniroma1.it/~navigli/pubs/ACL_2006_Navigli.pdf (15.05.2007).

- Orav, Heili 1998. Eesti keele direktiivverbide semantilise välja struktuur tesaurusena. – Magistritöö. Tartu Ülikool, üldkeeleteaduse õppetool, käsikiri.
- Orav, Heili; Vider, Kadri 2002. Kas tesaurus ja tekstid lähevad kasutuses kokku? – Tähendusepüüdja. Toim. R.Pajusalu, T.Hennoste. Tartu: Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 3, lk 297–303.
- Palmer, Martha; Ng, Tou Hwee; Dang, Hoa Trang. Evaluation of WSD Systems. .– Word Sense Disambiguation. Ed E.Agirre, P. Edmonds. Netherlands: Springer, Series: Text, Speech and Language Technology , Vol. 33, pp 75–106.
- Pedersen, Ted. Unsupervised Corpus-Based Methods 2006. – Word Sense Disambiguation. Ed E.Agirre, P. Edmonds. Netherlands: Springer, Series: Text, Speech and Language Technology , Vol. 33, pp 133–166.
- Peters, Wim; Peters, Ivonne; Vossen, Piek 1998. Automatic sense clustering in eurowordnet. In Proceedings of the 1st Conference on Language Resources and Evaluation (LREC). Granada, Spain;
<http://www.vossen.info/docs/1998/PetersPetersVossen.pdf> (15.05.2007).
- Peters, Wim; Peters, Ivonne 2000. Lexicalised Systematic Polysemy in WordNet. In: Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece; <http://gandalf.aksis.uib.no/non/lrec2000/pdf/148.pdf> (15.05.2007).
- Peters, Wim 2004. Detection and Characterization of Figurative Language. PhD thesis. University of Sheffield; http://www.dcs.shef.ac.uk/~wim/thesis_web.pdf (14.05.2007).
- Ravin, Yael; Leacock, Claudia (Eds) 2002. Polysemy : theoretical and computational approaches, pp 23.
- Resnik, Philip; Yarowsky, David 1997. A Perspective on Word Sense Disambiguation. Methods and Their Evaluation, Proceedings of ACL Siglex Workshop on Tagging Text with Lexical Semantics: Why, What and How? Washington, USA, pp 79–86.

- Resnik, Philip. WSD in NLP Applications 2006. – Word Sense Disambiguation. Ed E. Agirre, P. Edmonds. Netherlands: Springer, Series: Text, Speech and Language Technology, Vol. 33, pp 299–326.
- Saarikoski, Harri M.T 2005. Using Sense Categories and WSD System Scores to Generate Application-Specific sense inventories. In Proceedings of Human Language Technologies, Tallinn, Estonia, pp 179–182.
- Treumuth, Margus; Kikas, Taavet 2007. Word Sense Disambiguation; <http://math.ut.ee/~treumuth/NLP/semantics2.pdf> (23.05.2007).
- Tähenduspõhise keeletöötamise ressursid ja töövahendid eesti keele jaoks. 2003 – 2006. ETF grant nr 5534 (2003-2006) lõpparuanne (käsikiri). Tartu.
- Vider, Kadri; Muisnek, Kadri 2004. Sõnaliigituse kitsaskohad eesti keele arvutianalüüsis. Eesti Rakenduslingvistika ühingu aastaraamat 1(2004). Eesti Keele Sihtasutus, Tallinn 2005, lk 99–114.

Methods of Improving Word Sense Disambiguation for Estonian Language. Summary

The purpose of this M.A thesis is to propose methods how to make word sense disambiguation (WSD) a feasible task for Estonian language. Both automatic and manual WSD are kept in mind.

The first theoretical part gives an overview of general topics in WSD. It is described why do we need to discriminate between word senses and what kind of natural language processing applications need WSD. Also, the theoretical part shows the process of manual WSD and describes the different WSD Corpora in the world.

The second theoretical part is more specific – several methods of improving WSD task are proposed. Firstly, it is explained how the „one sense per one discourse” heuristics is applicable and considering this principle, all Estonian polysemous words are divided into two lists – the ones that follow „one sense per on discourse“, and the ones that doesn’t. One possibility to improve both automatic and manual WSD ist o examine the local context of target word in order to create rules. These rules should help to decide the right sense of the given target word. This is also the main practical tasks of this M.A thesis. Altogether 50 polysemous and frequent substantives are analyzed and 122 rules were created.

Another way towards WSD improvement is reducing the fine-grained sense inventory of the resources. Primarily it is done by clustering similar word senses in WorldNet. This could be done with many automatic methods, from which „Semantic Mirroring“ and systematic polysemy patterns are employed in the last, experimental part of the analyzes. This part gives some examples of manually clustered similar word senses.

Lisa 1 – Analüüsis viidatud käsitsiühestajate erimeelsuste failid

Esimeses tulbas on esitatud tähendused, mille vahel käsitsiühestajad kahtlesid (sõna+täh), teises tulbas on vastava erimeelsuse esinemise sagedus.

<i>sõna+täh</i>	<i>sagedus</i>
aeg;2,4	21
aeg;1,4	8
aeg;4,6	8
aeg;4,5	6
aeg;1,2	3
aeg;3,4	3
aeg;2,3	2
aeg;4,+1	2
aeg;1,5	1
aeg;1,6	1
aeg;2,6	1

<i>sõna+täh</i>	<i>sagedus</i>
asi;7,11	10
asi;3,11	7
asi;3,9	7
asi;3,7	5
asi;1,11	4
asi;3,8	4
asi;3,5	3
asi;7,9	3
asi;1,4	2
asi;1,5	2
asi;1,7	2
asi;10,11	2
asi;5,11	2
asi;1,9	1
asi;10,+1	1

<i>sõna+täh</i>	<i>sagedus</i>
mees;1,2	27
mees;1,3	3
mees;2,3	1

<i>sõna+täh</i>	<i>sagedus</i>
laps;1,2	17
laps;1,+1	1

<i>sõna+täh</i>	<i>sagedus</i>
naine;1,2	10
naine;1,3	9

<i>sõna+täh</i>	<i>sagedus</i>
aasta;1,2	18
aasta;2,4	6
aasta;2,3	1

<i>sõna+täh</i>	<i>sagedus</i>
kord;3,8	5
kord;3,+1	3
kord;2,3	2
kord;3,5	2
kord;1,7	1
kord;2,8	1
kord;3,7	1

<i>sõna+täh</i>	<i>sagedus</i>
käsi;1,5	14
käsi;1,4	12
käsi;1,+1	1
käsi;2,4	1
käsi;3,+1	1
käsi;4,+1	1
käsi;4,5	1

<i>sõna+täh</i>	<i>sagedus</i>
pea;1,2	11
pea;2,5	2
pea;1,+1	1
pea;1,1	1
pea;2,3	1
pea;3,5	1

<i>sõna+täh</i>	<i>sagedus</i>
mõte;1,2	11
mõte;1,5	10
mõte;3,4	3
mõte;1,3	2
mõte;1,4	2
mõte;2,3	2
mõte;2,5	2
mõte;4,5	2
mõte;3,5	1

<i>sõna+täh</i>	<i>sagedus</i>
uks;1,2	14
uks;2,+1	1

<i>sõna+täh</i>	<i>sagedus</i>
keel;2,3	8
keel;1,2	4
keel;2,4	4
keel;2,6	2
keel;1,3	1
keel;1,5	1
keel;2,5	1
keel;4,+1	1
keel;4,6	1
keel;6,+1	1

<i>sõna+täh</i>	<i>sagedus</i>
hää;2,3	6
hää;2,5	4
hää;1,2	1
hää;1,5	1
hää;2,4	1
hää;4,+1	1

<i>sõna+täh</i>	<i>sagedus</i>
kool;1,3	8
kool;1,2	2

<i>sõna+täh</i>	<i>sagedus</i>
töö;1,3	10
töö;1,6	5
töö;1,4	2
töö;1,2	1
töö;2,3	1
töö;3,6	1

<i>sõna+täh</i>	<i>sagedus</i>
koht;3,4	3
koht;1,4	2
koht;2,+1	2
koht;3,+1	2
koht;5,+1	2
koht;5,6	2
koht;1,7	1
koht;1,8	1
koht;2,6	1
koht;2,7	1
koht;2,8	1
koht;3,6	1
koht;4,5	1
koht;4,8	1
koht;6,8	1
koht;8,+1	1

<i>sõna+täh</i>	<i>sagedus</i>
pilk;1,3	7
pilk;2,3	7
pilk;2,+1	2
pilk;1,+1	1
pilk;1,2	1

<i>sõna+täh</i>	<i>sagedus</i>
tee;3,4	6
tee;1,2	2
tee;1,4	2
tee;1,3	1
tee;1,6	1
tee;1,7	1
tee;2,+1	1
tee;3,+1	1
tee;3,6	1
tee;4,6	1
tee;4,7	1
tee;5,7	1

<i>sõna+täh</i>	<i>sagedus</i>
tunne;2,3	10
tunne;1,2	6
tunne;1,3	1

Lisa 2 – Analüüsitud substantiivide esinemissagedused STÜ korpuses

Sõna	Tähendusnumber	Sagedus
aasta	2	94
aasta	1	30
aasta	4	5
aasta	2 või 3	2
aeg	4	163
aeg	2	42
aeg	5	9
aeg	1	5
algus	2	23
algus	3	4
algus	+1	1
asi	11	44
asi	7	15
asi	5	13
asi	3	13
asi	4	11
asi	9	9
asi	8	9
asi	1	6
asi	+1	5
asi	10	4
asi	6	1
elu	3	56
elu	7	31
elu	4	25
elu	2	6
elu	2 või 3	1
elu	6	1
elu	1	1
hing	4	11
hing	6	10
hing	5	3
hing	2	3

hing	3	1
hing	+1	1
hing	1	1
hää	2	42
hää	5	8
hää	3	4
hää	+1	2
hää	4	1
hää	1	1
hää	0	1
jutt	3	36
jutt	1	22
jutt	2	17
jutt	+1	7
jõud	3	10
jõud	2	9
jõud	4	5
jõud	+1	4
jõud	1	2
jõud	3 või 1	1
jõud	2 või 3	1
keel	2	45
keel	4	10
keel	3	2
keel	+1	2
keel	7	1
keel	5	1
keel	1	1
keha	3	34
keha	2	5
kell	4	20
kell	2	6
kell	+1	3

kell	3	1
kell	1	1
kodu	1	108
kodu	2	4
kodu	0	1
koht	4	21
koht	+1	18
koht	2	8
koht	3	5
koht	6	4
koht	1	3
koht	8	2
kool	1	43
kool	3	5
kool	+1	2
kord	3	92
kord	8	4
kord	+1	4
kord	2	3
kord	7	2
kord	1	2
kord	6	1
käsi	1	89
käsi	5	35
käsi	2	7
käsi	+1	7
laps	2	86
laps	1	57
laps	2 või 1	1
laps	+1	1
laud	1	43
laud	+1	7
laud	2	6
laud	3	3
lugu	2	21
lugu	1	19

lugu	+1	1
lõpp	7	9
lõpp	2	9
lõpp	5	8
lõpp	4	6
lõpp	1	3
lõpp	+1	2
lõpp	4 või 2	1
maa	7	34
maa	3	7
maa	8	4
maa	+1	4
maa	6	3
maa	2	3
maa	1	3
maa	4	2
maa ilm	3	12
maa ilm	+1	6
maa ilm	5	4
maa ilm	4	2
maa ilm	1	1
maja	1	61
maja	2	31
meel	3	10
meel	6	8
meel	4	5
meel	1	5
meel	5	3
meel	2	1
meel	+1	1
mees	1	225
mees	3	76
mees	2	36
mees	4	6
mõte	1	55
mõte	5	16
mõte	4	10

mõte	2	8
mõte	3	7
mõte	+1	4
naine	1	163
naine	3	151
naine	2	14
nägu	1	77
nägu	2	22
nägu	3	13
nägu	+1	3
olu_kord	4	22
olu_kord	4	1
olu_kord	2	1
olu_kord	1	1
osa	1	11
osa	4	6
osa	2	2
osa	+1	2
osa	9	1
osa	7	1
osa	5	1
osa	3	1
paber	4	14
paber	3	11
paber	1	6
paber	2	1
pea	1	70
pea	2	32
pea	+1	3
pea	2 või 5	1
pea	5	1
pea	4	1
pea	3	1
pilk	3	32
pilk	1	17
pilk	2	8
pilk	+1	7

poiss	1	77
poiss	2	9
poiss	+1	1
päev	4	37
päev	2	36
päev	5	24
päev	6	22
päev	3	3
päev	8	2
päev	4 või 6 või 5	1
rahu	2	16
rahu	1	3
rahu	+1	2
rahu	3	1
riie	1	14
riie	2	5
ruum	1	6
ruum	4	5
ruum	3	4
ruum	2	4
rõõm	1	15
rõõm	2	6
rõõm	3	1
samm	1	27
samm	3	6
samm	2	5
samm	+1	4
taevas	2	27
taevas	+1	4
taevas	1	3
tee	4	27
tee	3	7
tee	2	7
tee	1	7
tee	+1	5

tee	5	4
tee	6 või +1	1
tee	4 või 6	1
tee	6	1
tund	1	36
tund	+1	7
tund	2	3
tunne	2	29
tunne	3	9
tunne	1	5
töö	3	26
töö	1	16
töö	6	10
töö	4	5
töö	+1	5

töö	5	1
uks	1	73
uks	2	11
uni	1	10
uni	2	6
vari	2	8
vari	1	5
vari	+1	2
öö	1	29
öö	2	8
külg	3	8
külg	1	5
külg	2	1