TARTU ÜLIKOOL
FILOSOOFIA TEADUSKOND
Eesti ja üldkeeleteaduse instituut

Kristian Kankainen

Leksikograafi abivahend kakskeelse sõnastiku sõnavastete semantiliseks korrastamiseks paralleelkorpuse põhjal

Bakalaureusetöö

Juhendaja Arvi Tavast

Tartu 2013

Sisukord

1 Sissejuhatus

Käesolevas bakalaureusetöös käsitletakse leksikograafi abivahendit, mille peamine eesmärk on aidata semantiliselt korrastada automaatselt joondatud kakskeelse sõnastiku tõlkevastekandidaatide loendi sõnastikuks. Kirjeldatakse sellise abivahendi teoreetilisi probleeme ja visandatakse selle üldisi tööpõhimõtteid ning võetakse esimesi samme sellise süsteemi rakendamise poole, kusjuures defineeritakse esialgne andmestruktuur ja kaks lihtsat operatsiooni sellise struktuuri manipuleerimiseks.

Tõlkevastekandidaatide loend on masintõlke jaoks automaatselt loodud sõnapaariloend, mis sisaldab sõnu ja nende vasteid. Abivahend ei sea sisendile mingeid piiranguid vastete omavahelise vastavuse kohta.

Abivahend koosneb andmebaasist ja sellesse sisalduvaid andmeid analüüsivatest ja muutvatest funktsioonidest. Abivahend on mõeldud andmebaasi korrastamiseks ja ei käsitle andmebaasi algset ega lõplikku kuju (ega ka vahepealset seisundit) eraldi ega teistmoodi. Seega võib seda nimetada algebraliseks lähenemiseks sõnastiku koostamisele.

Andmed on salvestatud graafandmebaasis. Graaf on matemaatiline konstruktsioon, mille abil on võimalik kirjeldada objekte ja nendevahelisi seoseid. Objekte nimetatakse graafiteoorias tippudeks ning seoseid servadeks. Käesolev töö seisneb suuresti üldise graafi definitsiooni mõistmises, mille järgi graaf formaalselt koosneb kolmest komponendist: tipuhulgast, servahulgast ja intsidentsusfunktsioonist, kusjuures servahulk on tipuhulgast *a priori* sõltumatu. (vrd Buldas, Laud & Villemson 2003:11).

Käesoleva töö raames käsitletakse kõiki sõnu ja sõnavorme kui graafitippusid, kahe keele sõnade tõlkelist vastavust kui serva ning vaste keelesuunda kui intsidentsus-funktsiooni.

Töö eri osades käisitletakse sõnaga sarnaseid, ent põhimõtteliselt erinevaid üksusi. Sõna on defineeritud kolme eri metodoloogia järgi erinevalt, seetõttu tehakse siinses töös järgmist terminoloogilist vahet: 1) *sõna* ja *fraas* tähistab arvutuslingvistilist üksust, mis on tähtede jada ja võib sisaldada tühikuid; 2) *lekseem*iga tähistatakse üldlingvistilist tähenduslikku üksust ning 3) *lemma* varutakse abivahendis tähistama sama sõna eri sõnavormide komplekti.

Andmete tähenduste automaatseks analüüsimiseks on valitud Helge Dyviki semantilise peegli meetod, mis on kohandatud graafandmebaasis olevate sõnade

tähendusi dünaamiliselt liigendama.

Sõnade semantiline korrastamine abivahendiga on seega käsitsi töö, mille juures kasutajale esitatakse automaatselt grupeeritud samatähenduslikud sõnad, ning kasutaja ülesandeks on leida gruppidest need sõnad, mis tähenduse poolest erinevad teistest ja kustutada need. Vastavalt sellele, on abivahendi peamised funktsioonid mõistete jagamine ja ühitamine, mis on töö praktilises osas ka rakendatud graafandmebaasis.

Töö lähtekohaks on võetud sõnastikukoostamine sellise keelepaari puhul, millele on võimalik tekitada paralleelkorpus (tõlgitud veebilehed, programmid, e-raamatud jms) ja millel võib olla juba eelnevaid sõnastikke.

Töö struktuur on järgmine. Kuna töö lähtekoht eeldab, et tekstide paralleelistamine on lahendatud probleem, siis käsitletakse töö esimeses osas (pt 2) paralleeltekstide ekvivalentsuse ja komponeeritavuse mõisteid ning nendest tulenevaid probleeme lausest väiksemate üksuste (sõnade-fraaside) joondamisel ja võrreldakse (meta-)leksikograafia levinud ekvivalentsuse klassifikatsiooniga.

Töö teine osa (pt 3) hõlmab tõlkevastete semantilise korrastamise ja liigendamise aspekti. Eelkõige kirjeldatakse semantilise peegli tõlkelist semantikamudelit, mis on abivahendi keskseks osaks. Seejärel vaadatakse sarnasusi ja erinevusi ühe Eestis varem tehtud tööga. Seejärel (pt 3.4) näidatakse semantikamudeli implementatsiooni graafandmebaasis ja kirjeldatakse abivahendi üldisi tööpõhimõtteid. Abivahendi mõned töövõtted püütakse näitlikustada väikese korpuse peal (pt 3.5).

Viimases, 4. peatükis tehakse kokkuvõte ja väike arutelu edasise töö arenemisvõimalustest.

2 Leksikaalne joondamine ja ekvivalentsus

2.1 Paralleelkorpus ja paralleeltekst

Paralleeltekst[1] on tekst koos selle tõlke või tõlgetega ja paralleelkorpus tähistab niisiis kogumikku mitmest sellisest tekstist. Tekst ja selle tõlge võib esineda ühendatult mis tahes keeletasandil, ja sellisel puhul öeldakse, et tekst on joondatud vastaval keeletasandil, nt peatüki-, lause- või sõnatasandil.

1 Terminoloogiline märkus: terminit *paralleeltekst* kasutatakse tõlketeoorias teise tähendusega, kus see ei tähenda tõlget, vaid domeenisarnast teksti. Sellist teksti nimetatakse arvutuslingvistikas omakorda võrreldavaks tekstiks (*comparable text*) (Véronis 2000). Käesolevas töös järgitakse terminoloogiat nagu see on arvutuslingvistikas levinud, selliselt on see ka rootsi leksikograafias levinud (nt Svensén 2004:71).

Bo Svenséni leksikograafi käsiraamat (Svensén 2004:70) mainib paralleelkorpuste kasutamise kriitikaks leksikograafi seisukohalt, et kuna tõlkel on tendents olla seletavam kui originaaltekst, siis tuleks teha vahet ka retsiprookse paralleelkorpuse ja tavalise paralleelkorpuse vahel. Retsiprookne paralleelkorpus sisaldab võrdsel määral autentseid tekste, mis on tõlgitud suunal lähtekeelest sihtkeelde ning sihtkeelest lähtekeelde.

2.2 Paralleeltekstide joondamine

Joondamine tähendab kahe teksti sama tähendusega segmentide vastendamist ehk ühitamist. See toimub keele paradigmaatilisel küljel, paigutades joondatud ühikud nendele ühisele paradigmaatilisele teljele paralleelteksti mingi funktsiooni mõttes. Üldiselt on segmendisuuruseks valitud lause. Siinne töö ei süvene lausete joondamise metoodikasse ja problemaatikasse, vaid võtab lähtepunktiks juba olemasolevad vähemalt lausetasandil joondatud tekstid.

On tähtis märkida, et mistahes meetodiga joondatud laused on omavahelises *tõlkelises* vastavuses, s.o need on teksti raames sama kommunikatiivse funktsiooniga. Tõlkeline vastavus on konkreetse tõlkija teksti loomisel (s.o tõlkimisel) tehtud valikute tulemus, mis sõltub mitmest tema pragmaatilisest arusaamast tõlgitava teksti kohta (tekstiliik, teksti eesmärgid, teksti vastuvõtt, teksti kultuuriline kohandamine, jne). Sellisena ei eelda tõlkeline vastavus alati seda, et laused oleks ka *tähenduslikus* ehk semantilises vastavuses. (Hannesdottir 2001).

Sellisest kahe, s.o funktsionaalse ja semantilise ekvivalentsuse erinevusest tulenevad paljud probleemid kui joondatakse lausest väiksemal tasandil. Seda võib ehk näitlikustada naljade tõlkimisega – tõlkeekvivalendid võivad suurema segmendi tasandil küll mõlema keele lugejad sama hästi naerma panna, aga sõna-sõnaliselt ei pruugi need üldse samad naljad olla. Antud kohas on ka sobilik Martin Kay täheldus, et kuigi selliselt joondatud üksused võivad edendada tähtsaid väljavaateid edasiseks uurimiseks, võib selliste kirjete kaasamine kakskeelsesse sõnastikku põhjustada sõnastikukasutajale arusaamatusi ja frustratsiooni (Kay 2000:xviii). Kuigi Martin Kay täheldus on muidugi õige, ei ole see universaalne – sõnastikku valitud kirjed sõltuvad konkreetse sõnastiku funktsioonist. Mis ühte sõnastikku sobib, ei pruugi teise sobida ja vastupidi. Antud töö eesmärgiks on luua leksikograafile platvorm, kus just sellistele küsimustele vastata, mis sõnastikukasutajale näidata. Teisisõnu võimaldab abivahend paigutada need

arusaamatused ja frustratsioonid kuskile sõnastiku automaatse genereerija ja selle lõppkasutaja vahele.

2.3 Paralleeltekstide joondamine leksikaalsel tasandil

Joondamine leksikaalsel tasandil on ülaltoodud problemaatika tõttu raske. Olivier Kraif näitlikustab oma artiklites ((Kraif 2002; Kraif 2003)), et üldised tõlkeekvivalentsil põhinevad joondamismetodoloogiad viivad vasturääkivate tulemusteni leksikaalsel tasandil. Käesolev töö püüab läheneda probleemile mitmekülgselt, käsitledes antud peatükis kõigepealt ekvivalentsuse mõistet meta-leksikograafias, seejärel Kraifi tõlke-teoreetilist kriitikat tõlke komponeeritavuse kohta ja esitatakse tema leksikaalse vastavuse kontseptsioon. Peatüki lõpus näidatakse kuidas üks leksikaalse tasandi joondamisalgoritm töötab, mil määral see on mõjutatud tõlke komponeeritavuse mõistest ning pakutakse paar lihtsat edasiarendust, mistõttu algoritm võiks olla suuremaks abiks kavandatava abivahendi jaoks.

2.3.1 Leksikograafiline ekvivalentsus

Arlet Adamska-Sałaciak võtab väga mitmekülgselt kokku ekvivalentsuse mõiste leksikograafias, sünteesides tosinkonna teadlase vaatenurgad kokku. Ta nendib, et kuna tegu on väga laia mõistega, keskendub tema klassifikatsioon kakskeelse ekvivalentsuse mõistele, kuigi rangelt võttes kattub see ka ükskeelsete sünonüümide ekvivalentsiga (Adamska-Sałaciak 2010:387). Tema klassifikatsioon jaguneb neljaks erinevaks ekvivalentsuse klassiks (artiklis esitatud sünonüümsed nimetused on jäetud alles lugejale orienteerumiseks) (2010:397):

- kognitiivne ekvivalentsus (semantiline, süsteemiline, prototüüpiline, kontseptuaalne, dekontekstualiseeritud, mõisteline);
- seletav ekvivalentsus (deskriptiivne);
- tõlgitav ekvivalentsus (tekstisisestatav, tekstuaalne, kontekstuaalne);
- funktsionaalne ekvivalentsus (situatsiooniline, kommunikatiivne, diskursus, dünaamiline).

Kaks esimest klassi on n-ö interlingvaalsed, s.t vaatavad ekvivalentsust keelelisel-süsteemilisel tasandil (nii nagu seda teeb lingvistika), ja kaks viimast klassi on intertekstuaalsed, s.t vaatavad ekvivalentsust teksti tasandil (nii nagu teeb tõlketeooria ja

semantika). Niisiis hõlmab leksikograafia ekvivalentsuse mõiste mitut valdkonda.

Adamska-Sałaciaki järgi on euroopa traditsiooniline leksikograafia ajalooliselt tegelenud tekstisisese ekvivalentsusega, lisades tähtsatele tekstidele seletusi ja glosse. Alles palju hiljem on huvi liikunud keele kui süsteemi tasandile. Tänapäeva sõnastikud sisaldavad üldjuhul mõlema tasandi ekvivalente. (Adamska-Sałaciak 2010:chap. 2). Vaatame järgnevalt klassifikatsiooni nelja klassi.

Kognitiivseid ekvivalente iseloomustab väga üldine vastavus ja Adamska-Sałaciak peab neid kakskeelse sõnastiku tüüpilisteks kirjeteks. Selline vastavus sobib edasi andma lähtekeele märksõna üldist tähendust. Teistpidi võib vaadata seda prototüübiteooria järgi kui märksõna prototüüpseima tähenduse ülekandmist.

Kognitiivsete ekvivalentide klass on kõige altim sümmeetriale. Sümmeetria puhul on ekvivalentsus kahesuunaline, s.t on tõenäone, et selline tõlkevaste tõlgitakse tagasi lähtekeelde sama sõnaga (A→B→A). Samuti on need just kognitiivsed ekvivalendid, mida me saame kõige lihtsamini gradeerida tähenduskattuvuse järgi, ja sellise gradeerimise tõttu ei saa me alati loota sümmeetriale. (Adamska-Sałaciak 2010:397).

Kuigi seletavate ekvivalentide klass on sarnane kognitiivse klassiga, on nende vahel vahetegemine oluline. Adamska-Sałaciaki arvates seisneb vahe selles, et seletavaid ekvivalente leidub alati ja igas keeles, aga kognitiivne vaste ei pruugi olla ühes konkreetses (siht)keeles leksikaliseerunud ja sel puhul kognitiivne ekvivalent lihtsalt puudub. (Adamska-Sałaciak 2010:398).

Tõlgitava ja funktsionaalse klassi ekvivalendid on, nagu eelnevalt mainitud, teksti-põhised. Niisiis sõltuvad need lähtekeele üksuse kontekstist, seejuures ka laiendatud, pragmaatilisest kontekstist. Kui ekvivalent on sisestatav samasse konteksti nii, et tekib adekvaatne tõlge, kuulub ekvivalent tõlgitavasse klassi. Tihti aga, seletab autor, peab kasutama teisi strateegiaid selleks, et säilitada lähteteksti stilistilist, afektiivset jms tooni (seda fenomeni nimetab (Hannesdottir 2001:124) tõlketeoorias interferentsiks, ja peab seda ekvivalentsusemõiste tähtsaks osaks). Kasutades sihtkeeles teise grammatilise kategooriaga (sõnaliik jms) vastet, või mõnda sihtkeele idioomi, kuulub ekvivalent funktsionaalsesse klassi. Klassifikatsiooni järgi ei ole selge, kas funktsionaalne ekvivalents on eraldiseisev klass või kuulub see tõlgitava ekvivalentsi klassi alla, aga Adamska-Sałaciaki meelest piisab sõnatasemel vastavusest (*word-level correspondence*) nende eristamiseks. (Adamska-Sałaciak 2010:398–399).

Adamska-Sałaciak lisab oma artiklis, et ükski kakskeelne sõnastik ei saa sisaldada kõiki tõlgitava klassi ekvivalente, kuna on võimatu ette näha kõiki nende kontekste. See on loogiliselt muidugi õige, kui ideaaliks on loomuliku keele lõpmatuse idee, või ka siis, kui autor piirdub klassikalise trükitud sõnastikuga. Aga kui vaadata mingi kindla valdkonna keelt, nagu on näiteks arvuti graafilise kasutajaliidese keel, kui piiratud keelt, siis on (Tsepelina & Veskis 2010) kirjeldatud süsteem just seda laadi sõnastik, mis hõlmab terminite kõiki võimalikke kontekste.

Ekvivalentsuse suunalisuse aspekti järgi jagab Adamska-Sałaciak lihtsustades tõlketeooriad kahte poolusse, millest üks äärmus oletab loomuliku ekvivalentsuse eksisteerivat eelnevalt tõlkele, ja teine poolus arvab, et just tõlkeakt tekitab ekvivalentsuse. Teisiti öeldes tähendab see ühelt poolt (s.o esimese rühma teooriate pooldajate poolt), et tõlkija tegevus on loomuliku tõlkevaste *leidmine.* Teiselt pooluselt vaadatuna on tõlkija tegevus suunaline tõlke *tekitamine* (kusjuures suunda mõistetakse siin kui lähte- ja sihtkeele suunda). Vahetegemise tähtsus on suur, aga eeldab täpsustamist mis klassi ekvivalentsusega on tegemist (kognitiivne, seletav, tõlgitav või funktsionaalne ekvivalentsus). (Adamska-Sałaciak 2010:chap. 6). Adamska-Sałaciaki järgi on ainult kognitiivne klass oma loomult kahesuunaline, teised kolm konstrueeritakse *on-line* ja on palju komplekssemad (Adamska-Sałaciak 2010:400).

Adamska-Sałaciak toob enda klassifikatsiooni kriitikaks, et see piirdub leksikaalsete üksuste teatud tähenduste (*sense*) vahelise ekvivalentsusega. Adamska-Sałaciak nendib oma artiklis huvitavamate võimaluste olemasolu, aga tema järgi nõustuvad metaleksikograafid üldiselt, et leksikaalne ekvivalentsus valitseb leksikaalsete üksuste teatud tähenduste (*sense*) vahel. Aga leidub ka leksikograafe, kelle meelest sõnastikutähendused on vaid leksikograafilise analüüsi artefaktid. (Adamska-Sałaciak 2010:chap. 3).

Antud töö ei võta selles arutelus seisukohta, vaid sümpatiseerib nii selliste artefaktide kui ka nendevaheliste ekvivalentsuste tekkimisega.

2.3.2 Tõlke komponeeritavusest leksikaalse vastavuseni

Olles näidatud, missugused ekvivalentsussuhted võivad esineda kahe keele sõnatähenduse vahel, vaadatakse nüüd, kuidas lausest väiksemate üksuste joondamine neid käsitleb. Lugeja võiks pidada eelnevalt mainitud naljatõlkimise näidet orientiirina, kuidas ühel tasandil hästi joondatud segment võib väiksemaks segmenteeritult muutuda

kehtetuks ja halvaks joonduseks, lausa absurdiks. Probleemi põhjaks on segmenteerimiseks valitud suurus ja selliste segmentide ühitamiskriteerium.

Olivier Kraif (Kraif 2002:274) on toonud välja, et joondamismetodoologiad eeldavad üht spetsiifilist tõlketeoreetilist oletust, nimelt oletust tõlke kompositsionaalsusest. Tõlke kompositsionaalsus tähendab, et tõlkeekvivalent (s.t paralleeltekst, paralleellause) on komponeeritav, s.t koosneb väiksematest segmentidest, ja et terviktekst on nende väiksemate segmentide funktsioon. Teisiti vaadates tähendab see, et tõlketekst on segmenteeritav ja et mingil tasandil käituvad sellised segmendid kui teksti primitiivid.

Kraif kritiseerib sellise oletuse lingvistilist paikapidavust lausest väiksemate üksuste tasandil ja tema kriitika põhjaks on eeskätt sellise metodoloogia järgi saadud segmendi-üksuse iseloom – sellel puudub süntaktiline järjepidevus ja see on suvalise suurusega (võib olla nii sõna, sõnade jada, kui ka suurem üksus).

Lahenduseks pakub Kraif välja jagada joondamismetodoloogiad kaheks: üheks üldisemalt kehtivaks, mida nimetab maksimaalseks joondamiseks, ja üheks leksikaalsel tasandil kehtivaks, mida nimetab leksikaalseks vastavuseks. Nende erinevused on toodud välja tabelis 2.1. (Kraif 2002:286).

Kraifi leksikaalse vastavuse nõude olulisimaks punktiks on see, et joondatav üksus oleks keelespetsiifiline ühik (sõna, liitsõna, fraseologism vms). Antud töös arvestatakse sellist üksust kui sõnastikukirjet, mida lihtsustavalt nimetatakse märksõnaks. Töö kavandatav abivahend tõstab Kraifi leksikaalse vastavuse formaliseeringu raskused joondamisalgoritmilt inimleksikograafile. See ei tähenda, et joondusalgoritmi ei võiks muuta. Joondatava üksusega seotud probleem on ka lemmatiseerimine, mille kohta rohkem allpool.

Kraifi järgi on kompositsionaalsuse mõiste seotud ainult tõlkeprotsessi tulemusega, mitte tõlkimise kui tegevusega. See tähendab, et kompositsionaalsus on relatiivne ainult tõlketeksti kui produkti suhtes ja mitte kui tegevuse suhtes. Lõpuks järeldab ta, et tõlke kompositsionaalsus on see, mis määrab segmentide suuruse, milleks antud paralleelteksti on võimalik jagada, hoides alles segmentide joondatavuse ehk paralleelsuse. (Kraif 2002:274).

	Leksikaalne vastavus	Maksimaalne joondamine
Segmenteerimise kriteerium:	ükskeelne (see on keele-spetsiifiline) leksikaalse üksuse tasand	segmenteerimine sõltub tekstide struktuursest sarnasusest; põhineb nii tõlke kompositsionaalsusel kui ka maksimaalsusel (segmendid ei ole rohkem dekomponeeritavad)
Formaalne iseloomustus:	tihti üks-ühele seos leksikaalsete üksuste vahel; ülejäänud üksusi ei arvestata; on võimalik ka mitu-mitmele seosed.	kvaasi-bijektsioon, kvaasi-monotoonia lausest väiksemal tasandil
Segmentide süntaktiline iseloomustus:	leksikaalne üksus (sõna, liitsõna, fraseoloogia, jms)	puudub süntaktiline järjepidevus, võib olla sõna, fraas, lause, lõik
Ühitamiskriteerium:	denotatiivne identsus (esinemiskontekstipõhine)	tõlkeekvivalentsus

Tabel 2.1: Olivier Kraifi leksikaalset vastavust ja maksimaalset joondamist iseloomustavad jooned (Kraif 2002:286).

2.3.3 Joondamisalgoritm Anymalign

Kõigepealt tuleb nentida, et kui muud pole öeldud, tähendab selles peatükis termin *sõna* lihtsalt tühikute või lausepiiriga eraldatud tähtede jada, sest selliselt defineerituna on see tõepoolest arvutuslingvistikas levinud. Varume termile *leksikaalne üksus* sõna tähendust leksikograafilises-lingvistilises mõttes.

Anymalign (Lardilleux & Lepage 2009) on lausest väiksemate üksuste joondaja. Sellise joondaja eesmärgiks on koostada fraasitabel, mis on andmeesituse formaat, mida kasutavad mitmed statistilised masintõlkesüsteemid. Fraasitabel koosneb tõlkepaaridest (tõlkeekvivalentidest) koos skooriga, nt tõenäosushinnang ja sagedus. Lihtsalt öeldes kujutab fraasitabel endast sõnapaaride loendit.

Lardilleux jagab lausest väiksemate üksuste joondamismeetodid kahte suunda: tõenäosuslikud (*estimative approach*) ja seostavad (*associative approach*).

Esimese suuna algatas Brown et al. (1988. a) ja see põhineb paralleelkorpuse statistilise mudeli loomisel, mille juures parameetrid hinnatakse globaalse maksimeerimisega (s.o üle kõigi lausete üle terve korpuse). Eesmärgiks on jõuda parima hulga sõnadevaheliste seosteni (*alignment links*).

Teise suuna tõid välja Gale ja Church (1991. a) ja see põhineb mingil tõlkevaste-

kandidaatide genereerimise viisil ning selliste kandidaatide seoste tugevuse hindamisel statistiliste meetodite abil. Sellisena on hinnanguprotsess lokaalne, s.t igat segmenti käsitletakse lahus.

Populaarseim on neist matemaatiliselt rohkem põhjendatud tõenäosuslik lähenemine, ent selle parameetrite komplekssuse ja arvutuskeerukuse tõttu on püütud arendada seostava lähenemise mudeleid edasi. (Lardilleux, Lepage & Yvon 2011:190). Lardilleux on näidanud nende lähenemiste fraasitabelite üksteist täiendavat komplementaarsust (Luo, Lardilleux & Lepage 2011).

Anymalign käib seostava lähenemise alla, aga selle ajendiks on olnud mitmed teised puudused: vajadus rohkem kui kahe keele joondamisele (*multilingual alignment*), vajadus arvutuslikult skaleeritava algoritmi järgi (mida rohkem paralleelseid arvutusi, seda kiirem), metodoloogiline lihtsus (mis lihtsustab sellise vahendi integreerimist teistesse rakendustesse) (Lardilleux & Lepage 2009:214).

Kui teised statistilised meetodid toetuvad fenomenide kõrgetele sagedustele, käib Anymalign vastuvoolu, tuginedes madalatele ja ekstreemselt madalate sagedustele. Intuitiivseks praktikaks on kaua olnud korpuse suurendamine selleks, et tõsta madalate fenomenide sagedusi. Selline praktika on aga nõiaringiga – korpust suurendades lisanduvad uued harvaesinevad sõnad! Seni on metodoloogiliselt välditud harvaesinevaid sõnu neid väljafiltreerides, seda kritiseerides tuletab Lardilleux meelde, et suurem osa tekstist koosnebki harvaesinevatest sõnadest. Sellist väidet toetab ka Zipf'i jaotus. (Lardilleux, Lepage & Yvon 2011:191).

Hapax legomena (või lihtsalt hapaks) on nähtus, mille puhul sõna esineb korpuses vaid ühel ainsal korral. Näiteks on Shakespeare'i teostes keskmiselt 58% sõnadest hapaksid. Lardilleux järgi tulenevad harvaesinevad sõnad kahest keele omadusest: ühelt poolt sõnavara rikkusest ja teisalt selle sünteetilisuse astmest – seda sünteetilisem on keel, seda rohkem on selles sõnavorme (Lardilleux, Lepage & Yvon 2011:192).

Hapaks-sõnadel on üks hea omadus – need on teksti suhtes ühemõttelised. Kui sõna esineb tekstis ainult ühel korral, on sellel ainult üks funktsioon (teksti suhtes). Neid nimetab Lardilleux ideaalseteks joondusteks (*perfect alignment*). (Lardilleux & Lepage 2009:214).

Anymalign võtab kasu harvaesinevate sõnade nähtusest ja saavutab suurema sõnavara katvuse kui 2009. aastal täpsuse poolest parim olnud mudel (tõenäosuslik

mudel Giza++) (Lardilleux & Lepage 2009:218).

Anymaligni algoritm on valitud töös kirjeldatud abivahendi esimeseks proovitavaks allikaks just sõnavara suurema katvuse tõttu. Kuna harvaesinevate sõnade ideaalsete joonduste kasutamisel ei peaks tekkima muidu morfoloogiliselt rikaste keelte joondamisel levinud andmehõreduse probleem (nt Goldwater & McClosky 2005), ja võimaldab läheneda korpuse lemmatiseerimisele mõneti ebatraditsionaalsel moel (vt pt 3.4.3.1).

2.3.3.1 Algoritmi seletus

Illustratsioonis 2.1 on toodud välja Anymaligni joondamise põhimõtteline töö-printsiip. Selline seletus tugineb Anymaligni autori välja pakutud algoritmi seletavast lihtsustusest *minimalign.py*. Siinne kirjutaja püüab tuua välja algoritmide erinevusi, kus need põhimõtteliselt erinevad. Mõlemad algoritmid eeldavad, et sisestatud tekstid on juba joondatud lausetasandil.

```
 1:korrata mitu korda:
 2:  juhuslikult valitud alamkorpuse kohta:
 3:    tee iga sõna kohta loend selle jaotusest alamkorpuse lausetes
 4:    tee sõnaloendid sama jaotusega sõnadest
 5:
 6:    iga sellise sõnaloendi puhul:
 7:      alamkorpuse iga lähte- ja sihtkeele lause põhjal:
 8:        lähteKorr := need sõnaloendi sõnad, mis esinevad LK lauses
 9:                     järjestatud nii nagu need esinevad LK lauses
10:        lähteKomp := LK lause ülejäänud sõnad samuti järjestatud
11:        sihtKorr  := need sõnaloendi sõnad, mis esinevad SK lauses
12:                     järjestatud nii nagu need esinevad SK lauses
13:        sihtKomp  := SK lause ülejäänud sõnad samuti järjestatud
14:
15:        juhul kui leiti nii lähteKorr kui ka sihtKorr:
16:          salvesta paar joondusena ja lisa ta skoorile üks punkt
17:        juhul kui leiti nii lähteKomp kui ka sihtKomp:
18:          salvesta paar joondusena ja lisa ta skoorile üks punkt
19:
20:kui valmis:
21:  järjesta joondused skooride põhjal kahanevalt
22:  väljasta tulemused
```

Illustratsioon 2.1: Anymaligni algoritmi pseudokood (minimalign.py põhjal).

Nagu mainitud, kasutab Anymalign ära madalaid sagedusi, et saavutada n.n ideaalseid joondusi. Sellise efekti saavutab algoritm valides korpusest juhuslikke väikseid alamkorpusi (rida 2) ja neid analüüsides (rida 6). Korrates sellist protseduuri mitmeid kordi (rida 1) kaetakse tervet korpuse sõnavara, seetõttu on algoritm hästi skaleeritav ehk paralleelselt jooksutatav mitme arvuti protsessori peal; samas toob see ka kaasa omaduse, et programm on iga hetk lõpetatav – nt kui programmi on

jooksutatud kuni mingi piirarvu saavutamiseni, nt ajalise piiranguni või piisavalt madala uute joonduste suhteni sekundis, järjestatakse ja väljastatakse tulemused (rida 20).

Joondamisalgoritmi põhitöö seisneb sõnaloendite analüüsimises (rida 6), millest tulenebki Anymaligni üsna lihtne tõlkeekvivalentsuse kontseptsioon. Sõnade kattuvad jaotused võetakse kui ideaalne joondus mis on algoritmi ekvivalentsuse eelduseks.

Sõnaloendite (tegelikult sama jaotusega sõnade hulkade) analüüs seisneb selles, et alamkorpuse iga rida käiakse uuesti läbi, ja reas esinevad sõnad järjestatakse selle järgi, kuidas need konkreetses lauses esinesid (read 8 ja 11). Niiviisi saadakse maksimaalsed sõnade (süntagmaatiliselt järjestatud) „fraasid“ tekstilausest välja tõmmata ja salvestada muutujatesse 'lähteKorr' ja 'sihtKorr'. Seejärel salvestatakse ülejäänud lauses esinevad sõnad muutujatesse 'lähteKomp' ja 'sihtKomp' (read 10 ja 13). Matemaatiliselt võib seda vaadata kui 'lähteKorr' ja 'sihtKorr' komplementi, tõlketeoreetiliselt on see aga tõlke komponeeritavuse üks-ühele printsiibi otsene rakendamine. Algoritmi autor põhjendab sellist võtet, et on tõenäone, et lausete need osad on ka vastavuses (vrd Lardilleux, Yvon & Lepage 2012:281).

Järjestatud sõna-fraas salvestatakse tõlkeekvivalendiks juhul kui sõna-fraas leiti mõlemas keeles (rida 15 ja 17) (õiges Anymalign algoritmis ei käsitleta keeli lähte- ja sihtkeelena, ja keelte arv on piiramatu). Siinkohal võib märkida, et nii korrespondeeruvaid sõnu-fraase kui ka nende lausekomplemente käsitletakse sama laadi joondustena (read 16, 18 ja 21).

Joondatud paaride skoorisüsteem töötab sel viisil, et kui tõlkeekvivalent leitakse esimest korda, salvestatakse see joonduseks koos skooriga 1, järgnevateil kordadel, kui sama tõlkeekvivalent leitakse, lisatakse juba salvestatud joonduse skoorile 1 punkt. Hiljem, algoritmi lõpetamisel, arvutatakse nende põhjal ka tõlgete suunalised tõenäosused ning leksikaalsed kaalud (*lexical weight*).

Selline protseduur viiakse läbi alamkorpuse iga lause puhul, ja seejärel valitakse korpusest uuesti üks väike lausete ports, millele korratakse sama protseduur. Kuidas saavutatakse korpuse statistiliselt esinduslik katvus, on lähemalt seletatud (Lardilleux & Lepage 2009:chap. 3.2).

Kuna algoritm tegelikult joondab ekvivalentseid sagedusi juhuslikult valitud väiksetes lausete hulkades (alamkorpustes), võib küsida, missugused on need sõnad-fraasid mis sellistena leiduvad. Seda küsimust on Lardilleux käsitlenud (Lardilleux et al.

2009) ja leiab, et inglise-prantsuse keelepaari vahel on need eelkõige ühesõnalised komponendid (unigrammid). Algoritm lõhub pikemad sõnadejärjendid (mh püsiühendid ja kollokatsioonid) eraldi joondusteks. Töös pole uuritud, kuidas see mõjutab inglise-eesti suunal joondamist, ning pole seetõttu võimalik hinnata selle mõju abivahendile.

2.3.3.2 Algoritmi täiendused leksikograafi abivahendi jaoks

Algoritmiga leitud korrespondeeruvad ja komplementeeruvad joondused tuleks hoida eraldi. Kuna need on faktiliselt teistmoodi üksused (üks on vastavuses ja teine on oletatud vastavuses), võiks need abivahendi struktuuris ära märkida.

Teistsugune algoritmitäiendus oleks salvestada joondustele nende konkreetsed esinemiskohad korpuses. Kuna joonduse kvaliteet ei ole alati globaalselt laiendatav üle terve korpuse, oleks see viis kuidas tugevdada korpuse ja sellest tuletatud sõnastiku suhteid mitteüldistaval moel. Samuti annaks see võimaluse integreerida korpusetekstide morfoloogilise (ja muu lingvistilise) analüüsi iseseisvamal moel, nagu seda pooldab nt ISO standardiks saanud *Linguistic Annotation Framework*.

Mainitud täiendusi pole mahupiirangu tõttu töös rakendatud.

3 Sõnatähenduse ja sõnavara semantiline liigendamine

Sõnatähenduse semantilise liigendamise all mõistetakse mitmetähenduslike sõnade tähendusjaotuste leidmist. Arvutuslingvistikas tegeleb mitmetähenduslikke sõnadega kaks suunda. Esiteks sõnatähenduse ühestamine (*word sense disambiguation, WSD*), mille ülesandeks on mingil automaatsel moel valida kontekstis olevale sõnale selle õige tähendus, kasutades selleks eelnevalt fikseeritud tähenduste loendit. Sõnatähenduse liigendamise (*word sense induction, WSI*) puhul koostatakse seevastu loend sõna võimalikest tähendustest, vastavalt sõna esinemisjuhtumitele teksti(korpuse)s. Võib öelda, et *WSI* sisaldab *WSD* ülesannet, aga ei piirdu sellega. (vrd Nasiruddin 2013:196).

Selliselt defineerituna sõltub *WSD* lähenemine otse sellele ette antud fikseeritud tähenduste loendist. Sellist otsest sõltuvust on lihtne kritiseerida igal juhul, kui sellest jääb puudu, kas sõnade või tähenduste arvu poolest (nt uued kasutusvaldkonnad või terminoloogiad).

Sõnatähenduse liigendamisel võib tuua iseloomustavaks jooneks selle, et selgitatakse sõnatähenduste arv, mitte ei sõnastata tähenduse definitsiooni või kirjeldust – tähenduse „sisuks“ jääb viide mingi heuristika abil klasterdatud esinemisjuhtumite grupile (millele tuginedes võib inimene muidugi hiljem sõnastada tähenduse seletuse). Nasiruddini järgi ongi klasterdamine ainus joon, mis läbib kõiki *WSI* meetodeid. (Nasiruddin 2013:196).

Selliseid klasterdamisheuristikaid on mitmeid, nii ükskeelsetel tekstidel põhinevaid, kui ka paralleelkorpustel põhinevaid. Ükskeelsete tekstide sõnatähenduse liigendamise meetodid põhinevad sõna distributsionaalsel analüüsil, üldjuhul sõna lähiskonteksti kaudu (kollokatsioon laiemas, firthilikus mõistes). (Nasiruddin 2013:196). See on üks Ameerika strukturalismi klassikalisi teese.

Tõlkelised meetodid seevastu asendavad sõna distributsionaalse analüüsi selle sihtkeele vaste(te)ga (Nasiruddin 2013:198). Selliselt põhinevad need tähenduste leksikaliseerumisel eri keeltes, ja kui lähtekeele sõnale vastab sihtkeeles rohkem kui üks vaste, s.o divergeerib, on põhjust arvata lähtekeele sõna mitmetähenduslikkust. Üks selline meetod on Helge Dyviki semantilise peegli meetod, milles tähenduste liigendamine põhineb vastete ja vastete-vastete klasterdamisel.

Antud töös on ehk kasulik teha vahet *sõnatähenduse liigendamisel* ja *sõnavara*

semantilisel liigendamisel, milledest töö keskendub viimasele. Antud töös kasutatud Helge Dyviki semantilise peegli meetodist on selge, et sõnade tähendused sõltuvad teistest sõnadest ja nende tähendustest. Sellest tulenevalt ei liigendata töös kirjeldatud abivahendiga mitte ainult üksikute sõnade tähendusi (vrd sõnatähenduse liigendamise ülesanne), vaid korrastatakse tervet sõnavara. Ei ole selge, kas see tuleneb töös kasutatud meetodist, või on see üldse omane tõlkelistele sõnatähenduse liigendamise metodoloogiatele.

Kuna ühe sõna tõlkevastete-vasteid võib ka käsitleda kui selle parafraase, vaadatakse järgnevalt üht Eestis varem rakendatud parafraaside süsteemi ja tuuakse välja kuidas antud töös rakendatud meetod erineb sellest. Seejärel kirjeldatakse põhjalikumalt töös kasutatud meetodit ja viimasena kirjeldatakse selle rakendust tõlkevastekandidaatide korrastamise abivahendis.

3.1 Tõlkevasted ja parafraasid

Chris Callison-Burch on oma väitekirjas (Callison-Burch 2007) kirjeldanud automaatset paralleelkorpusepõhist parafraaside leidmise meetodit. Tema meetodi peamiseks eesmärgiks on laiendada statistilise masintõlke fraasitabelite sisu seal, kus ühe keelepaari vahel pole mingi sõna vastet leitud, aga teise keelepaari vahel on.

Meetodile leidub ka teisi rakendusi, ja Eestis on seda rakendanud Maarika Traat tekstikoostaja abivahendis „Automaatne parafraaside leidmine ning sõnade ja lühifraaside tõlkimine paralleelkorpuste abil“ (nt Traat 2010b; Traat 2010a).

Parafraasideks nimetab Callison-Burch (samas keeles) erinevaid viise väljendada sama sisu (Callison-Burch 2007:11). Formaalselt defineerib ta parafraasideks kõik fraasitabeli üksused, mis jagavad sama vastet teises keeles. Näiteks kui on joondatud sõnapaar $A_1 \rightarrow X$ ja leidub selline joondatud sõnapaar $A_2 \leftarrow X$ siis on A_1 ja A_2 teineteise parafraasid.

Kuigi antud töös rakendatud Helge Dyviki semantilise peegli meetodis ei nimetata otseselt sama osatähenduse all grupeeritud sõnu omavahel parafraasideks, on selge, et need on samatähenduslikud elemendid ehk parafraasid. Üldse põhineb töös kirjeldatud abivahend arusaamal, et WSI meetodite klastrid seovad samatähenduslikke komponente, ehk parafraase.

Meetodite kõige suurem metodoloogiline vahe seisneb selles, et Callison-Burch defineerib parafraasi osatähendusteks selle vasted, ja laseb kasutajal kitsendada

parafraaside hulga tähendust valides õige sõna-vaste paari (vrd Callison-Burch 2007:53–55; 78). Seevastu grupeerib semantilise peegli meetod parafraasi vasted osatähendusteks sõltuvalt veel nende vastete-vastete kattumisele. Tuginedes Adamska-Sałaciakile, võib öelda, et Chris Callison-Burchi meetod *tekitab* tõlkele tähenduse ning, et Dyviki semantilise peegli meetod *leiab* selle.

Niisugusena ei saa kasutada Callison-Burchi meetodi sõnatähenduse liigendamiseks, kuna efektiivselt leiaks see sõnadele sama palju tähendusi, kui nendel on leitud vasteid.

Dyviki semantilise peegli meetod ei deklareeri sõna vastete-vasteid (parafraase) omavahel sünonüümideks, enne nende polüseemsust välja selgitades. Sellises analüüsis tugineb ta John Lyonsi maksimaalse polüseemia kontseptsioonile, mida kirjeldatakse järgnevalt.

Callison-Burchi parafraaside meetodi sisendiks on fraasitabel lemmatiseeritud paralleelkorpusest, millest on joondatud pidevaid segmente. Ta mainib heaks edasiarendamise suunaks kasutada sellist leksikaalset joondamisalgoritmi, mis joondab mittepidevaid segmente (2007:36), seda on käesolev töö teinud Anymaligni abil.

3.2 Maksimaalse polüseemia printsiip

Mitmetähenduslikkuse puhul tehakse vahet polüseemia ja homonüümia vahel (esimesel juhul on tähendused omavahel seotud, teisel juhul ei ole). Juhtumite vahel võib piiritõmbamine osutuda väga raskeks.

John Lyons on toonud välja kaks radikaalset printsiipi, kuidas sellist piiritõmbamist üldse vältida. Üheks pooluseks on võimalus maksimeerida homonüümiat, mille puhul tuleb igale tähendusele eraldada omaette lekseem (vorm), ja teiseks pooluseks on maksimeerida polüseemiat, mille puhul lekseemi kõik tähendused koondatakse ühe ja sama vormi alla. (Lyons 1977:552–555).

Printsiipide erinevus seisneb niisiis selles, kas koondatakse lekseemid sama tähenduse alla või sama vormi alla. Kahest viisist pooldab Lyons selgelt viimast, kritiseerides esimese juures üldse tähenduste finiitse loendamise võimalust. (vrd Lyons 1977:554).

Lyons jagab veel ülaltoodud printsiibid kaheks, rangeks ja leebeks. Rangema puhul peab nt lekseemide sõnaliik, loendatavus jms kategooriad kattuma), leebema puhul loeb ainult vormiline ekvivalentsus (1977:555). Helge Dyvik käsitleb semantilise peegli meetodis lekseemi rangemas mõistes, arvestades ka sõnaliiki (Dyvik 1998:chap. 3.5).

Käesoleva töö abivahendis ei saa kõiki üksusi pidada lekseemideks ja on seega käsitletud lõdvemalt.

3.3 Helge Dyviki semantilise peegli meetod

Semantilise peegli meetod on formaalne semantikamudel, mis põhistab sõnade semantilise kirjelduse nende tõlkelistele tunnustele. Meetod on loodud automaatseks mõistelise sõnastiku (*thesaurus*) struktuuri deriveerimiseks (õigupoolest võimalikult automaatseks norra keele WordNeti koostamiseks), ja ei piirdu seega ainult sõnatähenduste semantilise liigendamisega. (Dyvik 1998).

Meetodi võib jagada neljaks osaks: lekseemide vastete leidmine, lekseemide tähendusjaotuste leidmine, sõnade grupeerimine tähendusväljadesse ja tesauruse kirjete loomine-sortimine. Käesolevas töös on piirdutud meetodi sõnade tähendusjaotuste leidmise osale.

Meetodi sisendiks on lekseemide ja nende tõlkevastete loendid, mida on siinse kirjutaja teada koostatud ainult käsitsi (joondamise ehk lekseemide vastavuse nõudeid on lähemalt kirjeldatud (Thunes 2003)). Kuna käesolevas töös kirjeldatud abivahendi sisend on fundamentaalselt erinev, käsitletakse seda allpool eraldi (pt 3.4.1).

Meetod on välja töötatud ja kasutatud Oslo Ülikooli inglise-norra paralleelkorpuse (Johansson, Ebeling & Oksefjell 2002) sõnade esmaseks tähendusjaotuste märgendamiseks ja ühestamiseks (Lyse 2003), mille järel on korpust olnud võimalik kasutada nii norra kui ka inglise keele statistilise *WSD* mudeli treeningmaterjalina.

Semantilise peegli meetod põhineb viiel eeldusel (Dyvik 1998):

1. semantiliselt lähedastel sõnadel on suuresti kattuvad tõlkevastete hulgad;
2. laia tähendusega sõnadel on rohkem tõlkevasteid kui kitsatähenduslikel sõnadel;
3. hüponüümide tõlkevastete hulgad sisalduvad nende hüperonüümide tõlkevastete hulkades (s.t on nende alamhulgad);
4. kontrastiivset mitmetähenduslikkust peetakse ajalooliseks viperuseks ja on üksikute sõnade idiosünkraatiline omadus, mis ei peaks kajastuma teistes keeltes;
5. eritähenduslikel sõnadel ei ole ühiseid tõlkevasteid, v.a kontrastiivse mitmetähenduslikkuse puhul, mida 4. eelduse järgi peaks esinema ainult ühe keele puhul.

Tõlgete kasutamist semantika teadmusbaasina õigustab Dyvik sellega, et tõlkimine on mitte-teoreetiline tegevus, kus tõlkija hindab väljendite semantilist sobivust osana oma tavalisest keelelisest tegevusest, heal juhul seostades nende tähendusi ka eri keelte kultuurilise taustaga. Samuti võimaldab see semantilise peegli meetodil siduda täis- ja osasünonüümia (ja hüperonüümia), ebamäärasuse (*vagueness*) ja kahemõttelisuse (*ambiguity*) mõisted sõltuvusse välisele baasile. (Dyvik 1998).

Semantilise peegli meetodi on võtnud kasutusele ja implementeerinud formaalse kontseptianalüüsi valdkonnas (*formal concept analysis*, *FCA*) Priss ja Old ning näidanud selle kasu kontseptiavastamisel (*conceptual exploration*) internetisõnastike põhjal (Priss & Old 2005). Kuigi sõnade vastete-vastete hulkasid on kasutanud mitmed teised uurijad, on nende kasutamine sõnade tähendusjaotuse liigendamiseks Prissi ja Oldi nentimisel ainuomane Dyviki meetodile (2005:2).

3.3.1 Sõnade tähendusjaotuse leidmine

Sõnade tähendusjaotuse liigendamine põhineb meetodi 4. ja 5. eeldusel ja grupeerib sõnade tõlkevasted gruppideks vastavalt nende vastete-vastetele. Dyvik mainib ka võimaluse määrata kattuvuse skaalana, ent ei formaliseeri seda. (Dyvik 1998).

3.3.1.1 Meetodi põhioperatsioon: tõlkekujutis

Meetodi põhitehteks on nn tõlkekujutis (*translation image*, *t-image*) mis on funktsioon mis tagastab lähtekeele sõna tõlkevasted sihtkeeles. Meetod defineerib järgnevad funktsioonid t-kujutise abil (def 1—5, Dyvik 1998):

- Esimest järku t-kujutis (*first t-image*) on lähtekeele sõna tõlkevasted sihtkeeles. Näitlikustamaks, ütleme, mingi lähtekeele sõnal *a* on kolm tõlkevastet, *x*, *y* ja *z*.
- Pööratud esimest järku t-kujutis (*inverted first t-image*) on omakorda eelmise funktsiooni vastete t-kujutiste hulgad. Teisiti sõnastades on see lähtekeelsed vastete-vasted, grupeeritud sihtkeelsete vastete põhjal.
  Näitlikustamaks, kui sihtkeele sõnadel *x*, *y* ja *z* on vastavalt lähtekeele tõlkevasted *a* ja *b*, *a*, *b* ja *c* ning *a* ja *d*, siis pööratud esimest järku t-kujutis on hulkade hulk lähtekeele sõnadega $\{\{a,b\},\{a,b,c\},\{a,d\},\}$. On selge, et kõikide selliste alamhulkade ühisosaks on vähemalt *a*.
- Teist järku t-kujutise (*second t-image*) operatsioon on keerulisem. Olgu U inverteeritud t-kujutise hulkade summa. Teist järku t-kujutis on U-{a}

elementide t-kujutised (U-st jäetakse välja algne sõna). Veel lisatakse need sihtkeele elemendid, milledel on ainsaks vasteks algne sõna *a* (muidu jääksid need arvestamata). Teist järku t-kujutis on niisiis hulkade hulk, mis sisaldab sihtkeele sõnu. Selline „kolmanda ringi“ tõlkevastete hulk on suur ja võib sisaldada väga palju sõnu, mis ei ole algse sõnaga semantiliselt otseselt seotud.

- Kitsendatud teist järku t-kujutis (*restricted second t-image*) on teist järku t-kujutise hulkade ühisosad esimest järku t-kujutisega, see tähendab, et hulkade elemendid on kitsendatud sisaldama ainult algse sõna tõlkeid (s.o esimest järku t-kujutis!). Selle funktsiooni mõte on grupeerida esialgse t-kujutise sõnad vastavalt nende inverteeritud t-kujutise vastetele.
- Sõna tähendusjaotus on seega defineeritav kui kitsendatud teist järku t-kujutise mitte-kattuvate hulgad. Mitte-kattuvate hulgad saadakse, kui liidetakse omavahel kokku kõik teist järku t-kujutise hulgad, mis ühe või rohkem elemendi puhul omavahel kattuvad. Selle funktsiooniga grupeeritakse lähtekeele sõna omavahel parafraseeritavad vasted. Mitte-kattuvate hulkade arv on järelikult sama, mis sõna tähenduste arv.

Niisiis jaguneks eelmiselt näitlikustatud sõna *a* vasted kaheks tähendusek, mille ühele tähendusele vastaks niisiis tõlkevasted *x* ja *y*, ja teisele tähendusele *z*. On lihtne järeldada, et *x* ja *y* ning *z* tõlkevasted lähtekeeles parafraseeriksid vastavalt neid tähendusi.

Kui sõna tähendusjaotus on leitud, saab Dyvik käsitleda sõna-tähendus paare märkidena, mille peal opereerimisel on semantilise peegli meetodil võimalik defineerida komplekssemaid operatsioone (grupeerida sõnu tähendusväljadesse, ammutada leksikaalseid seoseid ja tesaurusestruktuuri). Käesoleva töö piiratud mahu tõttu, ei ole neid operatsioone töös rakendatud. Siiski näib siin hea koht tutvustada meetodi laiemaid võimalusi eesti keeles.

3.3.2 Sõnade grupeerimine tähendusväljadesse

Dyvik formaliseerib sõna-tähenduspaarid kui järjestatud paarid, kusjuures esimene element on sõna ja teine element on selle tähenduse tõlkevasted. Vastavalt meie näitele oleks meie märgid siis järgnevad: <a, {x, y}> ja <a, {z}>.

Märkide tähendused nimetab Dyvik lihtsalt A_1 ja A_2, niisiis oleks eelmise

notatsiooniga samaväärne ka <a, A_1> ja <a, A_2>. (Dyvik 1998:chap. 3.7).

Märke on seega sama palju kui tähendusi (vrd maksimaalne homonüümia).

T-kujutise operatsiooni rakendatakse märkidele sarnaselt kui sõnade puhul. Mainida tuleb seda, et märgi esimest järku t-kujutis sisaldab sellele märgile vastava osatähenduse vasted (nt A_1 t-kujutis on {x, y} ja A_2 t-kujutis on {z}).

Semantilise peegli meetod defineerib kaks märki kuuluma samasse tähendusvälja juhul, kui need jagavad üht või rohkem tähendusi teise keele suhtes, s.o kui mõlemad lähtekeele märgid esinevad ühe sihtkeele märgi esimest järku t-kujutises. Samuti kuuluvad kaks märki samasse tähendusvälja, kui on olemas selline sihtkeele märkide ahel, kus märkide t-kujutised omavahel kattuvad, ja lähtekeele üks sõna kuulub sihtkeele ühe märgi t-kujutisse ja lähtekeele teine sõna sellise ahela mõne teise sihtkeele märgi t-kujutisse. (def 7, Dyvik 1998).

Tähendusväljade kattuvuse põhjal on Dyvik ka näidanud üht viisi, kuidas on võimalik leida seotuid sõnu ehk tõlkevastekandidaate (Dyvik 2003).

3.3.3 Dyviki leksikaalsed seosed ja tesaurusestruktuur

Märkide seotus tähendusväljadesse võimaldab nende tähenduste jaoks indutseerida tunnuseid ja koostada sõnadele semantilise kirjelduse, mille järel on võimalik määrata sõnade leksikaalseid suhteid, kasutades meetodi kolm esimest eeldust (vt lk 18): nt on üks märk teise suhtes hüperonüüm, sel juhul, kui kõik selle märgi tunnused sisalduvad teise tunnuste hulgas. Meetod võimaldab siiski tuvastada ainult semantiliselt sarnaseid suhteid (sünonüümia, hüperonüümia ja hüponüümia), mitte nt antonüüme. (rohkem nt Dyvik 1998; Dyvik 2005).

3.4 Graafipõhine tõlkevastekandidaatide semantilise korrastamise abivahend

Käesoleva peatüki eesmärgiks on visandada abivahend, mis aitab leksikograafil määrata sõnade-vastete valikut kakskeelse sõnaraamatu koostamisel paralleelkorpuse põhjal. Tuuakse välja sellise programmi tööprintsiibid, töötava graafi-mudeli ning funktsioonid. Tööle ei ole veel realiseeritud visuaalset kasutajaliidest, aga graafi struktuuri püütakse hoida võimalikult ikoonilisena, et tulevikus võimaldada sellele otse visuaalset vaadet. Töö autor arvab, et selline sõnade-vastete graafi visuaalne näitamine

on kasutajale intuitiivne[2].

Lähtekohaks on võetud situatsioon, kus leksikograafile on kättesaadaval paralleelkorpuse sõnatasandil joondatud fraasitabelid või sõnapaariloendid. Kuna selline sõnadevastete loendi tekstiline esitus on väga pikk ja raskesti ülevaadatav, pakub abivahend sellise loendi tihedalt seotud andmekogumina ehk kompaktsemat, ülevaatlikumat ja semantilise organiseerituse poolest loogilisemat vaadet. Võimaldades mitu omavahel seotut vaadet samadele andmetele, on abivahend kindlasti saanud inspiratsiooni Douglas Engelbarti kasutajaliidestealasest tööst[3].

Sõnapaariloendi semantiline korrastamine klassikaliseks kakskeelseks sõnastikuks tundub intuitiivselt käesoleva töö autorile koosnevat järgmistest etappidest:

- leida sõna tähendused
- leida tähendustele vasted
- leida märksõna ja vaste kasutuse-tähenduse erinevused (ja kirjeldada need)

Kahe esimese punkti puhul võib argumenteerida, kas leksikograaf mõtleb lähtuvalt sõnast (semasioloogiliselt) või lähtuvalt tähendusest (onomasioloogiliselt). Käesolev töö ei taha võtta seisukohta sellises diskussioonis, vaid püüab luua keskkonda, kus need vaatenurgad on n-ö ühe medali kaks külge. Abivahendi keskkond püüab niisiis luua vaadet, kus tähenduse ja tähendusele vastavate vastete vahel ei tehta vahet.

Eelnevalt on näidatud, kuidas sõnade vasteid on võimalik automaatselt grupeerida osatähenduste kaupa. Seega on võimalik luua sõnapaariloenditele vaade, mis põhineb osatähenduste näitamisel. Kuna ühe osatähenduse sees peavad kõik elemendid olema teine-teisega semantiliselt seotud (olema parafraasid), on kasutajal lihtne märgata vigu ehk teine-teisega mitte kokkukuuluvaid elemente.

Sellisel struktuuril on eelis ka lihtsalt joondusvigade leidmisel. Joondamisel võib tekkida kahte tüüpi vigu – sõna on varustatud vale vastega või puudub õige vaste, ja võib arvata, et puuduvate vastete minimeerimisega tõuseb valede vastete arv. Graafilises esitluses on vead hõlpsamini ülesleitavad, kuna joondus on kahe elemendi vaheline seos. Seetõttu tuleb abivahendis joondusviga esile vähemalt kahel korral (mõlema vaste puhul). Kui struktuur on seotum, siis tuleb viga rohkemalgi korral esile, vead võivad ulatuda vastete-vasteteni. Sellist fenomeni võib nimetada vigade nakatuseks (vrd Roark & Charniak 1998:1112; Widdows & Dorow 2002:3). Võimalik on ka, et selliste

2 Võrdle nt Visual Thesaurus (http://www.visualthesaurus.com/)
3 Vaata nt http://www.dougengelbart.org/firsts/dougs-1968-demo.html

veakobarate leidmine on mingi päringuheuristika abil automatiseeritav, aga sellist võimalust pole töös uuritud.

Samuti võimaldab selline vaade lihtsamini liikuda seotud sõnade vahel.

3.4.1 Lemmatiseerimine

Sõnavara korrastamine kirjeldatud abivahendiga põhineb niisiis semantilise peegli meetodiga automaatselt leitud osatähenduste hulkade käsitsi korrigeerimisel. Üks fundamentaalne erinevus mida tuleb märkida, on see, et semantilise peegli sisendiks on lekseemid, mistõttu on osatähendusteks jagumine ennustatav. Abivahendi sisendiks ei ole tähenduslikud sõnad, vaid on ainult vastavusse joondatud tekstikatkendid (vrd sõne). Sellest osutub esimeseks probleemiks sõnavormide üldistamine lemmadeks. Seda tehakse abivahendis kahel moel. Esiteks seotakse sama lemma sõnavormid kokku kasutades morfoloogilist analüüsi, teisalt seob abivahendi kasutaja sõnavorme kokku käsitsi, kasutades semantilise peegli meetodi tulemusi selleks, et veenduda õigete seoste olemasolus. Kuna leksikograafias võib sõnavorm kanda teist tähendust kui selle lemma, on antud töös valitud säilitada sõnavormid.

Sellise põhimõttelise vahe tõttu ei ole võimalik vaadata käesolevat tööd kui hinnangut semantilise peegli meetodile kasutades automaatse joondaja sisendit.

3.4.2 Andmebaasi kirjeldus

Dyvik on kirjeldanud semantilise peegli meetodi sõnatähenduse liigendamise osa hulgateoreetiliselt ja realiseerinud need programmeerimiskeeles Lisp kasutades liste (Dyvik 2002). Sellisena on kõik andmed vaja laadida arvuti töömälusse ja ei ole seega väga laiendatav – mida rohkem andmeid, seda rohkem töömälu on tarvis.

Käesoleva töö andmestruktuuriks on valitud graaf. Viimaste aastate jooksul on selline andmestruktuur kogunud populaarsust (nt semantiline veeb ja sotsiaalmeedia) ja vastavaid andmebaasisüsteeme on hakanud ilmuma riiulitootena.

Andmebaasisüsteemiks on valitud Neo Technology mõõdukalt vaba litsentsiga graafandmebaasisüsteemi *Neo4j Community Version*. Kuna töö kontseptuaalselt ei sõltu ühestki Neo4j-le spetsiifilisest omadusest, näib selle andmebaasisüsteemi lähem kirjeldus ülearune. Mainida võiks siinkohal siiski suuremaid erinevusi graafandmebaaside ja arvutuslingvistikas muidu populaarsete relatsioonandmebaaside vahel. Graafandmebaasidel on võimalik protseduurilised päringud, s.o kirjeldada mitte ainult *mida*

otsitakse (deklaratiivne) vaid ka *kuidas* seda leitakse. Graafandmebaasi andmestruktuuri suurim eelis on selle paindlikkus – nii salvestamisel (andmeid võib puududa või suvaliselt juurde lisada, st ei pea täitma tabeli kõiki ettenähtud veergusid), kui ka otsimisel (kuna relatsioone ei pea ette defineerima, ei piira andmete struktuur *per se* selle võimalike päringute hulka).

Kuna nii andmestruktuuriks kui ka andmebaasiks on graaf, peetakse neid edaspidi sünonüümsetena.

3.4.3 Andmeallika lisamine

Andmestruktuuri paindlikkus võimaldab kasutada mitu informatsiooniallikat paralleelselt. Informatsiooniallikaid saab graafi mõistes olla kahte erinevat liiki, sellised, mis lisavad graafi tippusid ja sellised, mis lisavad servasid. Üldjuhul lisab üks allikas mõlemaid elemente, aga nendest tähtsam on servade lisandumine.

Infoallika lisamine peab andmete suhtes alati olema idempotentne, see tähendab, et midagi ei kustutata ning olemasolevaid elemente ei ole võimalik topelt lisada (sama atribuutidega element ei saa graafis esineda rohkem kui ühel korral).

Tüüpilised andmeallikad oleks nt eri joondamisalgoritmide abil leitud ja eri paralleeltekstidest joondatud fraasitabelid. Andmeallika sõnad-vasted lisatakse tippudena koos vastavate atribuutidega (keel) ja nendevaheline joondus lisatakse servana (atribuudid vastavalt joondusalgoritmile). Andmed märgistatakse ka päritoluallika nimetuse ja kirjeldusega.

Teistsugune, ent sisuliselt sama informatsioonistruktuuriga andmeallikavorm on sõnastik. Siinkohal tuleb ilmsiks tarvidus anda allikatele eri hinnanguid nende usaldusväärsuse kohta. Näiteks võib sõnastikust pärit informatsioon kaaluda rohkem kui automaatse joondajaalgoritmist pärit. Ka võib oletada, et käsitsi lisatud informatsioon on kõrge kaaluga. Kaalusid peab olema võimalik muuta globaalselt.

3.4.3.1 Morfoloogilise informatsiooni lisamine

Morfoloogilise informatsiooni (sh lemmatiseerimine) käsitlemine informatsiooni-allika erijuhtumina võib esialgu tunduda võõras. Eesti keele jaoks on olemas kolm eri automaatset morfoloogiaanalüsaatorit ja nendele lisanduvad veel ühestajad. Ka on eesti keele kohta näidatud, et ka spetsialistide käsitsi märgendamisel ei saavutata 100% märgendajatevahelist kooskõla (Kaalep et al. 2000:624).

Sõnavormi ja selle lemma ühendab graafis suunaline serv `has_lemma`, mille atribuutideks on mh info kasutatud lemmatiseerija kohta. Selline struktuur võimaldab kasutada mitmeid lemmatiseerijaid paralleelselt. Kui joondamisalgoritm oleks säilitanud viited esinemisasukohtadele tekstis, oleks võimalik ka pidev uuendamine, nt kui ühestajate täpsus tõuseb.

Kuna praegune töö põhineb maksimaalsel polüseemial, ei ole lemmatiseerimise täpsus niivõrd oluline, oluline on sõnade üldistamine siduvateks elementideks.

3.4.4 T-kujutise operatsioon koos lemmatiseerimisega

Lemmatiseerimine on funktsioon, mis antud sõnele (graafitipule) tagastab kõik selle võimalike lemmade sõnavormid graafis (graafitipud):

$$Lemma(sõne) \rightarrow sõne_1, sõne_2, \ldots, sõne_N$$

Niiviisi ei eksisteeri lemma graafis *per se*, vaid on t-kujutise omadus:

$$Tkujutis(Lemma(sõne))$$

Lemmatiseerimisega t-kujutise operatsioon töötab samal viisil kui tavaline t-kujutis, ainult, et ühe sõna tõlkevastete asemel liidetakse sõnavormi kõigi lemmade sõnavormide tõlkevasted kokku. Vastavalt lemmatiseerimisele kustutatakse teist järku t-kujutise saamisel inverteeritud t-kujutise ühendist kõik sama lemma sõnavormid.

3.4.5 Tööoperatsioonid

Järgnevalt tuuakse välja abivahendi tööoperatsioonid. Nendest on töö praktilises osas (pt 3.5.1.3) ainult rakendatud kahte esimest (mõistete jagamine ja liitmine).

Graafis on sõnad-fraasid ja lemmad salvestatud tippudena ning nende vahel on suunaline serv `has_lemma`. Töös on tõlkevastavus lihtsustatud mittesuunaliseks servaks `translates`.

Semantika poolest on graafis servad tähendusrikkamad kui tipud, ja niikaua kui opereeritakse servade peal, põhjustatakse ainult lokaalseid muutusi graafis. Semantilise peegli osatähendused ei ole graafis representeeritud tippudena, vaid on kogu aeg dünaamiliselt-virtuaalselt esitatud. Seetõttu kajastavad need graafi muutumist „globaalselt“ servade kustutamisel-lisamisel.

Tipu kustutamisega kaasneb kõikide sellega seotud servade kustutamine. Seega on see komplekssem tegu ja ei ole soovitatav ettenähtud abivahendis. Üldiselt võiks tipu saada kustutada ainult siis, kui sel puuduvad servad, või on kõik selle servad ainult

seotud endaga (silmused).

3.4.5.1 Mõiste jagamine

Juhul kui sõnapaaride loendis on selline viga, mis ühendab kaks mitte-parafraseeritavat sõna omavahel, tekib analüüsitaval sõnal mõiste, mille sees need mõlemad sõnad esinevad.

Sellisel juhul on abivahendiga lihtne markeerida need kaks (või rohkem) sõna, mis ei peaks olema omavahel seotud, ja lasta abivahendil näidata nendevahelised seosed korpusest. Seda teeb abivahend graafiteoreetilise tehtega, mis tagastab kõigi sõnade tippudevahelised ahelad ehk teed. Sellise loendi pealt on kasutajal võimalik leida vigane serv üles, ja kustutada see. Vigane serv võib olla nii `has_lemma`, kui ka `translate`. Kui kõik vigased servad on kustutatud, eralduvad sõnad oma mõistete alla.

3.4.5.2 Vaste ühendamine lemmaga (mõiste liitmine)

Kui ühe sõna osatähenduste loetelus esineb üks vaste eraldi sellega seotud mõistest, valib kasutaja vaste ja lisab selle ja mõistest valitud lemma vahele serva `has_lemma`.

3.4.5.3 Mõiste kehtestamine

Praktilise sõnastiku koostamise seisukohalt peaks kavandatav süsteem võimaldama peale sõnade ja vastete ka mõistete salvestamist graafis. Mõistet tuleb siinkohal mõista vaid tipuna, mis seob omavahel vastavusse kahe keele sõnu, umbes nagu klassikalise sõnastiku märksõna. Seega pakuks see kasutajale võimalust väljendada, et tema on kvalitatiivselt hinnanud selliste sõnade vastavuse, ning et kui ülejäänud sõnavara semantilisel korrastamisel asjaolu muutub, on ta huvitatud võrrelda erinevust selle ja dünaamiliselt genereeritud mõiste vahel. Sarnast, mitu-mitmele võimaldavat mõistete sõnastikustruktuuri on pakutud välja värvinimede puhul (Tavast et al. 2013).

Operatsiooni jaoks on vaja kavandada sünkroniseerimismehhanism, mis võrdleks selliste fikseeritud mõistete ja semantilise peegli meetodi poolt leitud virtuaalsete osatähenduste vahekorda, ning teavitaks kasutajat ebakõladest.

3.4.5.4 Mõiste varustamine seletusega

Abivahend peaks võimaldama lisada mõistete seletusi mõlema keele jaoks. Loogiliselt sisaldub selles operatsioonis ka mõiste kehtestamine.

3.4.5.5 Tõlkevaste lisamine

Abivahend peaks võimaldama ka tõlkevastete ühendamist serva `translates` lisamisega. Kui tõlge üldse puudub, jääb selle lisamine ülal mainitud lemma lisamise alla.

3.4.5.6 Servade kustutamine

Servade kustutamise põhjuseks on peetud ülal mainitud mõiste jagamine ning jääb seetõttu selle juhtumi alla. Muidugi peaks abivahendi detailsem vaade võimaldama kasutajale kõiki graafi elemente manipuleerima.

3.4.5.7 Seotud tähendusväljade näitamine

Sõnade kuulumine tähendusväljadesse on määratud semantilise peegli meetodi 7. definitsioonis. Abivahendi sellise vaate juures peaks võimaldama kasutajal valida tähendusvälja „laiust“ dünaamiliselt.

3.4.5.8 Tõlkevastekandidaatide pakkumine

Selleks leidub kirjanduses mitmeid algoritme, Helge Dyvik on kirjeldanud üht mis põhineb semantilise peegli tähendusväljadel (Dyvik 2003). Sellise vaate puhul ei peaks abivahend koormama leksikograafi visuaalset vaadet, vaid järjestama kandidaadid seotuse järgi kahanevalt. Seda võiks ka värviintensiivsusega näidata.

3.5 Väikese korpuse sõnavara korrastamise näide

Selles peatükis on näitlikustatud ülal kirjeldatud abivahendi funktsionaalsust ühe väikese korpuse sõnavara semantilise liigendamise puhul.

Korpuseks valiti vabalt kättesaadaval olevad KDE arvutitöölauakeskkonna eesti keele tõlked. KDE (*K Desktop Environment*) on avatud lähtekoodiga vabatarkvaraline graafiline kasutajaliides paljude lisaprogrammidega, mille eestindajad on kaua aega aktiivsed olnud ja mille väljakujunenud tõlkepoeetika on olnud suhtluses suurema vabatarkvara eestindamise kommuuniga ning võimalik, et ka laienenud sellele (nt Tepper 2009; Põder 2010; Lepik 2009; Põder & Tepper 2011; Põder 2012).

3.5.1 Korpuse koostamine

Väga sarnast korpust on koostanud (Tsepelina & Veskis 2010), ja on muidugi väga küsitav, miks käesolev töö ei võtnud seda aluseks. Seda tehti kahel alusel, esiteks on

nende sõnad-vasted lemmatiseeritud, teisalt taheti proovida teistsugust joondamisalgoritmi ja töökäiku. Edaspidine koostöö kindlasti ei ole välistatud.

3.5.1.1 Algkorpus joondatud sõnatasandil (korpus A)

Korpuse koostamiseks laaditi kõigepealt alla eesti keelse KDE kõik tõlkefailid (GNU gettext'i formaadis *po*-failid, mille kohta rohkem nt (Lepik 2009:4)). Selliste failide tõlkeelemendid sisaldavad nii sihtkeele (eesti) kui ka lähtekeele (inglise) tekstid. Seejärel liideti kõikide failide tõlkeelemendid kokku üheks suureks *po*-failiks (kasutades programmi GNU gettext msgcat), millest kustutati kõik korduvad elemendid (programmiga GNU gettext msguniq). Lõpuks vormistati suur *po*-fail ümber joondajale sobivaks tabeliformaadiks (programmiga Okapi tikal -2tbl). Sellesse tabelisse oli jäänud 25 444 tõlkeelementi.

Siinkohal on oluline märkida, et tõlkefaile ei lausestatud, seega sisaldas koostatud paralleelkorpus erineva suurusega tõlkeelemente, lühimad olles ühesõnalised menüükäsud ning pikemad mitmelauselised abitekstid. Tõlkeelementide semantilist vastavust peeti olulisemaks kui tõlkeelementide lauselist suurust. See oli vale otsustus kasutatud joondusalgoritmi suhtes, mis eeldab lausestatud sisendit. Liiga suurte tõlkeelementide joondamisel peaks Anymalign leidma väga pikki sõnakatkendeid, ja võib-olla seetõttu ei ole sellised sõnad tekitanud abivahendis nähtavaid efekte.

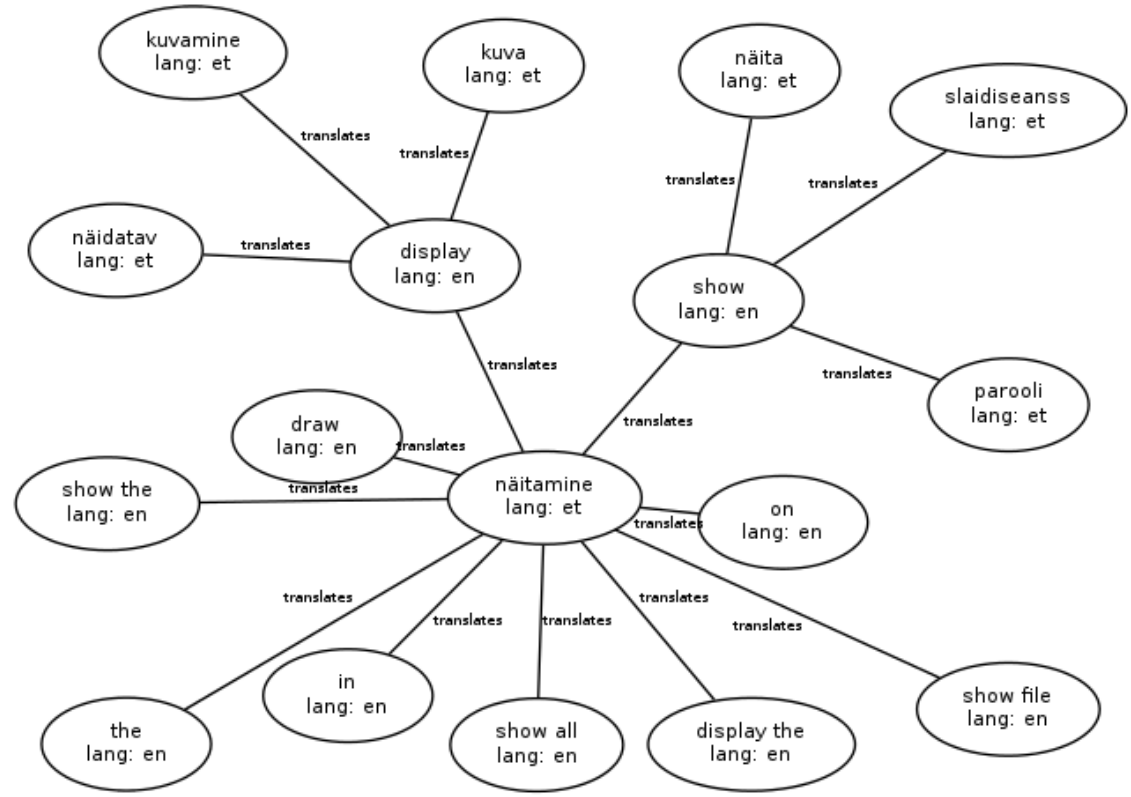
Paralleelkorpus joondati lausetasandil jooksutades Anymaligni algoritmi kaheksa tundi. Joondamisest saadud tabel oli suur ja sisaldas üle 7,5 miljoni tõlkepaari. Enne tõlkepaaride sisestamist graafandmebaasi, vormistati kõik tekst läbivalt väiketähtedeks ja puhastati igasugustest märgenduste märkidest (html, xml jne) ja kustutati tüüpilised kirjavahemärgid. Graafi sisestati lõpuks ainult tähtedest koosnevad termid, mille joondussagedus oli kõrgem kui 200. Joondussagedus on arv, mis näitab mitmel korral algoritm joondas sama tõlkepaari. Selliseid tõlkepaare oli 66 533 tükki. Selline sageduse piirmäära valimine oli täiesti *ad hoc*.

Kuigi Anymalignist saadud vasteteloend sisaldab mõlema keelesuuna tõenäosusi, lihtsustati tõlkemudel semantilise peegli meetodi jaoks sümmeetriliseks. Korrektsem oleks olnud säilitada suund, aga selline võimalus jäeti edasiarendamisvõimaluseks.

Korpus A puhul oli märgata üllatav terminoloogiline ühtlus, väga väike osa sõnedest oli seotud rohkem kui ühe vastega. Samuti tähendab see väga vähest kontrastiivset polüseemiat sõnavormide tasandil. Ei ole uuritud kas tegu on Anymaligni joondus-

algoritmi eripäraga, kuigi selle heaks küljeks peaks olema just harvaesinevate vastete leidmine. Asjaolu võib selgitada teadmine, et KDE on suuresti ühe inimese tõlgitud ja toimetatud.

Korpus A graafistruktuuri on püütud näitlikustada illustratsioonis 3.1.



Illustratsioon 3.1: Lemmatiseerimata joondustabeli struktuur näitlikustatud graafis.

3.5.1.2 Lemmatiseeritud korpus (korpus B)

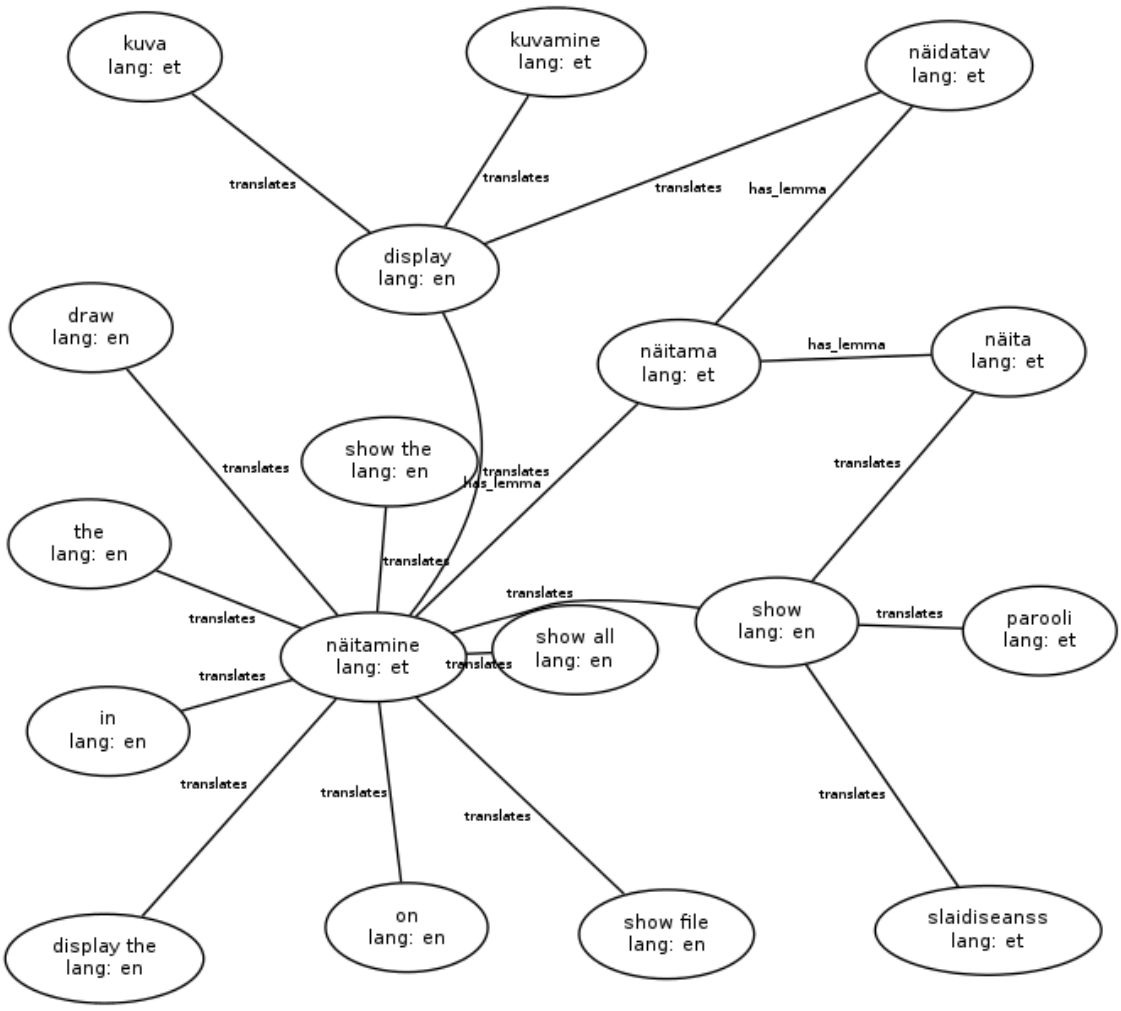
Morfoloogilist informatsiooni lisati ainult eesti keele jaoks. Seda tehti simuleerimaks sõnastikuprojekti andmekaevanduse *worst-case* stsenaariumi, kus ei ole kindel, et teise joonduskeele jaoks on olemas piisavad keeletehnoloogilised ressursid. Võib aga lisada, et inglise keele morfoloogia lihtsus siiski aitas sellisele otsustusele kaasa. Morfoloogia analüüsimiseks kasutati EKI analüüsimootorit[4], aga nagu eelnevalt mainitud, võimaldab töö graafistruktuur kasutada mitmeid analüüse paralleelselt, ja ei sõltu otseselt ühestki analüüsimootorist.

Morfoloogilisest informatsioonist on tööle hädavajalik ainult lemmatiseerimine, kuna selle abil üldistatakse sõnavormide seotust maksimaalse polüseemia põhimõttel.

4 http://artur.eki.ee/morf/

Morfoloogilise info sisestamisel lisati igale sõnavormile serv `has_lemma` sõnavormi ja lemma vahel. Juhul kui lemmat ei esinenud graafis, lisati lemmale oma tipu. Juhul kui sõnavorm langes kokku lemmaga, lisati ainult serv sõnavormilt iseendale (s.o silmus).

Korpus B graafistruktuuri on püütud näitlikustada illustratsioonis 3.2.



Illustratsioon 3.2: Lemmatiseeritud joondustabeli struktuur näitlikustatud graafis

3.5.1.3 Abivahendi rakendamine sõnavastete korrastamisel (korpus C)

Siinses peatükis on püütud rakendada ülal kirjeldatud abivahendi funktsionaalsust graafandmebaasis olevate sõnade mõistete automaatseks liigendamiseks, ning rakendada kaks kirjeldatud operatsiooni (mõiste jagamine ja mõistete liitmine) nende mõistete korrastamiseks. Märkida tuleb programmi visuaalse kasutajaliidese puudumine, operatsioonid on läbiviidud programmaatiliselt ja genereeritud tähendused

on näitlikustatud tabelitena.

Näitlikustamaks abivahendi esialgset tööpõhimõtet, valiti üks mitmetähenduslik sõna, mille vasteid korrastada. Sõnaks valiti eesti keele *kaart*.

Programmi esimene, puhtalt joondatud korpusepõhine, osatähenduste liigendamine on näidatud tabelis 3.1. Abivahendis rakendatud semantilise peegli meetod on genereerinud 21 mõistet. Suure mõistete arvu peamiseks põhjuseks on inglise keele vastete lemmatiseerimise puudumine. Lemmatiseerimise töö peab nüüd tegema kasutaja. Ta liidab kõik '*tab*'-iga mõisted kokku, lisades nende ja tipu '*tab*' vahele serv `has_lemma`. Sarnaselt seob ta omavahel kokku '*map*' mõisted.

Kuigi vasted nagu '*close*', '*to show on*', '*accounts*', '*of*', '*the*', '*in*', '*the accounts*', '*advanced*', '*general*', '*the comment*' ja '*the variables*' võib-olla ütlevad palju korpusega kirjeldatud maailma kohta (nt on kaart tähenduses 'tab' tõepoolest asi, mida saab KDEs sulgeda ja mille peal on võimalik midagi näidata), kustutab ta need ära. Kuna kustutatud vastete hulgas olid ka kõik vasted, mis sidusid kaks erinevat mõistet kokku üheks suureks mõisteks (kaart20), jagunesid selle tõttu mõiste õigesti kaheks mõisteks (vastavalt '*tab*' ja '*map*' tähendused). Uus genereeritud tähendusjaotus on näidatud tabelis 3.2.

KDE korpuse sõnastik (Tsepelina & Veskis 2010) annab sõnale 'kaart' kaks vastet: '*tab*' ja '*card*' (nt helikaart, graafikakaart) aga ei sisalda seevastu '*map*' vastet. Abivahendi andmebaasis on vastavad sõnad ('*soundcard*', '*graphics card*') olemas, ja veel teisigi ('*vcard*' ehk visiitkaart). Neid pole aga lemmatiseerija suutnud siduda 'kaart' lemmaga. Samuti on puudu kokkulangevaid tõlkevasteid inverteeritud ja teist järku t-kujutises, ent on võimalik, et need vasted asuvad lähedates tähendusväljades.

Mõiste number	**Mõistele vastavad vasted**
Kaart1	*view tab*
Kaart2	*the accounts*
Kaart3	*tab to*
Kaart4	*map of*
Kaart5	*the tab*
Kaart6	*the comment*
Kaart7	*the variables*
Kaart8	*the tabs*

Kaart9	*folder tabs*
Kaart10	*default map*
Kaart11	*center map on*
Kaart12	*current tab*
Kaart13	*of a tab*
Kaart14	*in map*
Kaart15	*to show on*
Kaart16	*the current map*
Kaart17	*from map*
Kaart18	*tabbed browsing*
Kaart19	*tab, close, of, the, in, page, advanced, general, maps, map*
Kaart20	*tabs*
Kaart21	*accounts*

Tabel 3.1: Sõna 'kaart' esimene tähendusjaotus genereeritud joondatud korpuse põhjal.

Mõiste number	**Mõistele vastavad vasted**
Kaart1	*view tab, tab to, the tab, the tabs, folder tabs, current tab, of a tab*
Kaart2	*maps, map of, default map, center map on, in map, the current map, from map*

Tabel 3.2: Sõna 'kaart' tähendusjaotus genereeritud pärast semantilist korrastamist.

4 Kokkuvõte

Töös alguses kirjeldati ühe joondusalgoritmi tööprintsiipi ning näidati kuidas see põhineb ühel tõlketeoreetilisel oletusel mida on kritiseeritud. Toodi välja kaks algoritmi parandust töös kavandatud leksikograafilise abivahendi töö jaoks, kuigi töö piiratud mahu tõttu ei rakendatud neid.

Seejärel tutvustati Helge Dyviki semantilise peegli meetodit kui üht tõlkelist sõnatähenduse liigendamisalgoritmi. Semantilise peegli meetod on kavandatud abivahendi keskseks osaks.

Kavandati leksikograafi abivahendi andmestruktuuri ja üldisi tööpõhimõtteid, mille juures olulisimaks komponendiks oli võimaldada sõnapaariloenditele vaadet, mis ei tee vahet sõnade tähenduste ja vastete vahel, kuna grupeerib vasted automaatselt samatähenduslikeks gruppideks. Kavandatud abivahendi põhiline töökäik seisneks leksikograafi selliste gruppide käsitsi paranduste sisseviimisel, mille tagajärjed kajastuksid talle automaatselt tagasi.

Abivahendi tööpõhimõtte testimiseks koostati väike sõnatasandil joondatud korpus, mis salvestati graafandmebaasi. Seejärel toodi välja semantilise peegli sisendi ja koostatud korpuse põhimõtteline erinevus, ja muudeti graafandmebaasi struktuuri võimaldama lemmatiseerimist (ka mitu erinevat paralleelselt).

Semantilise peegli sõnatähenduse semantilise liigendajaalgoritm kohandati töötama graafistruktuuri peal, ja töö lõpus näidati kuidas leksikograaf saab korrastada ühe sõna osatähendusi graafistruktuuris kahe operatsiooni abil.

Töö edasiarendamise võimalused on laiad. Eelkõige pakub huvi semantilise peegli meetodi teiste osade teostamine graafistruktuuris. See võimaldaks kasutajal eksploratiivselt suurendada-vähendada vaadeldava sõna semantilist välja ning leida sellega seotuid sõnu.

Sõnapaariloendi graafistruktuuri lähem uurimine oleks võimalik lisades sellesse semantilise peegli inverteeritud ja teist järku t-kujutiste servad, see võimaldaks graafi struktuurilisi omadusi uurida graafiteoreetiliste meetmete abil. Samuti võiks nende servade olemasolu lihtsustada tähendusväljade arvutamist.

5 A lexicographer's tool for semantically organizing a parallel corpora derived bilingual dictionary. Summary

The work is divided into three parts, where the first part makes an overview of the problems occurring when aligning parallel texts for smaller units than the textual sentence. The second and main part defines a graph structure for these lexical alignments and implements a word sense induction method for the same graph structure. Also a tool is sketched for organizing the graph structure semantically. The third and final part of the work demonstrates two simple functions for partial organization of the graph structure into groups containing semantically similar items.

The first part shows problems of aligning units smaller than the text sentence, where the problem is argued to be the lack of linguistic equivalence between all the aligned lexical units. Alongside an overview of lexicographic equivalence is given. Two modifications for the used aligning algorithm are proposed but not implemented.

The work's main focus is on defining a graph structure for lexically aligned parallel texts and sketching a tool for organizing the structure into semantically similar groups (e.g paraphrases).

Helge Dyvik's semantical mirror method has been chosen as the 'motor' for the semantical analysis of the lexical alignments directly in the graph structure. This gives the sketched tool the ability to dynamically reflect the groups of translations corresponding to the meanings of the analyzed word, thus enabling the user to review the sense distinctions and spot out semantically non-similar translations in groups where there should be only semantically similar ones.

Because the graph structure contains raw lexical alignments produced by Anymalign on unlemmatized texts, it creates a type-token problem in which the user has to define a lemma unit for tokens of the same type. The graph structure is expanded to hold lemma information data from automatic morphological analysis software. This structure is defined in such a way that it can hold multiple data sources in parallel.

In the works practical part, two simple graph-editing functions are demonstrated to split and to merge a polysemous word's meanings into a structure more resembling a bilingual dictionary.

6 Kirjandus

Adamska-Sałaciak, Arleta. 2010. Examining Equivalence. *International Journal of Lexicography* 23(4). 387–409.

Buldas, Ahto, Peeter Laud & Jan Villemson. 2003. *Graafid*. Tartu: Tartu Ülikool, matemaatika-informaatikateaduskond, arvutiteaduse instituut.

Callison-Burch, Chris. 2007. *Paraphrasing and Translation*.

Dyvik, Helge. 1998. A Translational Basis for Semantics. In Signe Oksefjell (ed.), *Corpora and Crosslinguistic Research: Theory, Method and Case Studies*, 51–86. Rodopi. http://hf.uib.no/i/LiLi/SLF/Dyvik/transem.html (21 February, 2013).

Dyvik, Helge. 2002. Semantic Mirrors in Medley Interlisp User's Guide. http://www.hf.uib.no/i/LiLi/SLF/ans/Dyvik/mirrorsguide.html (19 April, 2013).

Dyvik, Helge. 2003. Heuristisk utvidelse av t-bilder fra korpus. http://www.hf.uib.no/i/LiLi/SLF/Dyvik/T-utvidelse.pdf (14 April, 2013).

Dyvik, Helge. 2005. Translations as a Semantic Knowledge Source. *The Second Baltic Conference on Human Language Technologies: Proceedings, April 4-5, 2005, Tallinn, Estonia*, 27–38. Tallinn: Institute of Cybernetics. (18 February, 2013).

Goldwater, Sharon & David McClosky. 2005. Improving statistical MT through morphological analysis. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, 676–683. (25 June, 2013).

Hannesdottir, Anna Helga. 2001. Ekvivalent och ekvivalent – det beror på vad man menar. *Festskrift till Martin Gellerstam den 15 oktober 2001: Gäller stam, suffix och ord*(29). (Meijerbergs arkiv för svensk ordforskning). 122–136. (27 September, 2012).

Johansson, Stig, J Ebeling & Signe Oksefjell. 2002. *English-Norwegian Parallel Corpus: Manual*. 1999/2002 ed. University of Bergen: Department of British and American Studies. http://www.hf.uio.no/ilos/tjenester/kunnskap/sprak/omc/enpc/ENPCmanual.html (2 April, 2013).

Kaalep, Heiki-Jaan, Kadri Muischnek, Kaili Müürisep, Andriela Rääbis & Külli Habicht. 2000. Kas tegelik tekst allub eesti keele morfoloogilistele kirjeldustele? Eesti kirjakeele testkorpuse morfosüntaktilise märgendamise kogemusest. *Keel ja Kirjandus*(9). 623–633.

Kay, Martin. 2000. Preface. *Parallel text processing: alignment and use of translation corpora*, xv–xx. (Text, Speech and Language Technology vol. 13). Dordrecht [etc.]: Kluwer Academic Publishers.

Kraif, Olivier. 2002. Translation Alignment and Lexical Correspondences: A Methodological Reflection. In Bengt (ed. Altenberg & Sylviane (ed. Granger (eds.), *Lexis in Contrast: Corpus-Based Approaches*, 271–289. (Studies in Corpus Linguistics (StCoL): 7). Amsterdam, Netherlands: Benjamins.

Kraif, Olivier. 2003. From translational data to contrastive knowledge: Using bi-text for bilingual lexicons extraction. *International Journal of Corpus Linguistics* 8(1). 1–29.

Lardilleux, Adrien, Jonathan Chevelu, Yves Lepage, Ghislain Putois & Julien Gosme. 2009. Lexicons or phrase tables? An investigation in sampling-based multilingual alignment. (Proceedings of the 3rd Workshop on Example-Based Machine Translation). 45–52.

Lardilleux, Adrien & Yves Lepage. 2009. Sampling-based multilingual alignment. *Proceedings of Recent Advances in Natural Language Processing (RANLP 2009)*, 214–218. Borovets, Bulgaria. (1 October, 2012).
Lardilleux, Adrien, Yves Lepage & François Yvon. 2011. The Contribution of Low Frequencies to Multilingual Sub-sentential Alignment: a Differential Associative Approach. *International Journal of Advanced Intelligence* 3(2). 189–217. (27 November, 2012).
Lardilleux, Adrien, François Yvon & Yves Lepage. 2012. Hierarchical Sub-sentential Alignment with Anymalign. *Proceedings of the 16th annual conference of the European Association for Machine Translation (EAMT 2012)*, 279–286. Trento, Italie. (26 November, 2012).
Lepik, Sander. 2009. Vaba tarkvara tõlkimine Eestis Mozilla Firefoxi näitel. Informaatika Instituut: Tallinna Ülikool Seminaritöö. http://sander85.com/download/vaba_tarkvara_tolkimine_eestis.pdf (7 August, 2013).
Luo, Juan, Adrien Lardilleux & Yves Lepage. 2011. Exploring N-grams Distribution for Sampling-based Alignment. (5th Language & Technology Conference (LTC'11)).
Lyons, John. 1977. *Semantics*. . Vol. 2. Cambridge [etc.]: Cambridge University Press.
Lyse, Gunn Inger. 2003. Fra speilmetoden til automatisk ekstrahering av et betydningstagget korpus for WSD-formål. (1 July, 2013).
Nasiruddin, Mohammed. 2013. A State of the Art of Word Sense Induction: A Way Towards Word Sense Disambiguation for Under-Resourced Languages.
Priss, Uta & L. John Old. 2005. Conceptual Exploration of Semantic Mirrors. *Formal Concept Analysis: Third International Conference, ICFCA 2005*. Springer Verlag.
Põder, Märt. 2010. à la "mingi sõnade värk": Tarkvaratõlke juhenditest. http://tuiutama.blogspot.com/2010/10/p-margin-bottom-0.html (8 August, 2013).
Põder, Märt. 2012. Tarkvara tõlkimine (juhend) – Pingviini Viki. http://viki.pingviin.org/Tarkvara_t%C3%B5lkimine_(juhend)#T.C3.B5lkimise_printsiibid_ja_viljakad_t.C3.B6.C3.B6v.C3.B5tted (7 August, 2013).
Põder, Märt & Hasso Tepper. 2011. Stiilijuhend tarkvara tõlkimiseks – Pingviini Viki. http://viki.pingviin.org/Stiilijuhend_tarkvara_t%C3%B5lkimiseks (7 August, 2013).
Roark, Brian & Eugene Charniak. 1998. *Noun-Phrase Co-Occurrence Statistics for Semi-Automatic Semantic Lexicon Construction*.
Svensén, Bo. 2004. *Handbok i lexikografi : ordböcker och ordboksarbete i teori och praktik*. Stockholm :: Norstedts akademiska förlag.
Tavast, Arvi, Mari Uusküla, Kelly Parker & Urmas Sutrop. 2013. Using probabilistic conceptual graphs for representing colour terms in dictionaries. In Maurizio Rossi (ed.), *Color and Colorimetry Multidisciplinary Contributions*, xx–xx. Rimini: Maggioli Editore. [ilmumas]
Tepper, Hasso. 2009. eesti:kde:stiilijuhend [hasso.linux.ee].
Thunes, Martha. 2003. Ekserpering av oversettelseskorrespondanser fra parallelltekst.
Traat, Maarika. 2010a. *Automaatne parafraaside leidja*. http://ats.cs.ut.ee/parafraasid/ (13 May, 2013).
Traat, Maarika. 2010b. Automaatne parafraaside leidmine ning sõnade ja lühifraaside

tõlkimine paralleelkorpuste abil. http://www.keeletehnoloogia.ee/projektid/automaatne-parafraaside-leidmine-ning-sonade-ja (13 May, 2013).
Tsepelina, Katrin & Kaarel Veskis. 2010. Paralleelkorpuspõhine tõlkeabisüsteem internetis. *Keel ja Kirjandus*. 820–835.
Véronis, Jean (ed.). 2000. *Parallel text processing: alignment and use of translation corpora*. (Text, Speech and Language Technology vol. 13). Dordrecht [etc.]: Kluwer Academic Publishers.
Widdows, Dominic & Beate Dorow. 2002. A Graph Model for Unsupervised Lexical Acquisition. 1093–1099. (23 July, 2013).

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina __Kristian Juha Ismo Kankainen_____
(*autori nimi*)
(sünnikuupäev: __1984.09.11_____)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose _Leksikograafi abivahend kakskeelse sõnastiku sõnavastete semantiliseks korrastamiseks paralleelkorpuse põhjal,
(*lõputöö pealkiri*)

mille juhendaja on _Arvi Tavast_____,
(*juhendaja nimi*)

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

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

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

Tartus/Tallinnas/Narvas/Pärnus/Viljandis, _2013.08.28___ (*kuupäev*)