Emakeele Seltsi aastaraamat 60 (2014), 171–194
doi:10.3176/esa60.09

LEKSIKOSEMANTILISTE SUHETE HÄGUSUS EESTI WORDNETIS

HEILI ORAV, SIRLI ZUPPING,
KADRI VARE

Annotatsioon. Artiklis antakse ülevaade Eesti Wordnetis sagedamini kasutatavatest semantilistest suhetest, milleks on sünonüümia, hüperonüümia, rollisuhted, meronüümia ning hägussuhted. Kirjutises keskendutakse mõistete semantiliste suhete määramise ja kirjeldamise probleemidele, millele Eesti Wordneti kui mõistelise andmebaasi koostamisel lahendusi otsitakse. Tuuakse näiteid suhete kohta, mis on Eesti Wordnetis ebaühtlased ning mis vajaksid kohandamist eesti keelele. Samuti otsitakse lahendust entsüklopeediliste ja üldkeelsete tähenduste sidumiseks. Käsitluse laiemaks taustaks on maailma eri keelte *wordnet*'ide ühendamine, mis aitaks luua mitmekeelseid keeletehnoloogilisi ressursse.
Võtmesõnad: leksikaalne semantika, arvutileksikoloogia, keeleressursid, arvutilingvistika, eesti keel

1. Sissejuhatus

Üle veerandsaja aasta tagasi sai alguse leksikosemantiliste andmebaaside loomine, mida kutsutakse üldnimega *wordnet*. Selle ideeline põhi on leksikaalsete üksuste võrgustik, kus seotus tuleneb teatud fikseeritud suhete valikust. Algidee oli luua sõnavõrgustiku tüüpi mentaalne leksikon ehk mudel selle kohta, kuidas sõnad meie peas asetuvad ja kuidas need omavahel seotud on (vt Fellbaum 1998). Kirjeldatud leksikon oli mõeldud esmalt psühholoogide ja keeleteadlaste jaoks, nende uurimistulemuste peegeldamiseks, kuid praeguseks on *wordnet* pigem keeletehnoloogide tähelepanu all olev ressurss (Kilgarriff 2000).

Wordnet'i tööpõhimõte on peegeldada maailmateadmisi semantilise võrgustiku kaudu, mille sõlmpunktid on sünohulgad ja ühendavad kaared nende vahel on fundamentaalsed semantilised suhted. Martha Palmer (2009) on osutanud, et järjest rohkem otsitakse viise, kuidas annaks omavahel kombineerida eri vaatenurki. Lisaks on tehtud mitmeid uurimusi

selle kohta, kuidas klassifitseerida semantilisi suhteid (Evens 1988; Bejar jt 1991). Semantilises võrgustikus asuvad semantilised suhted teevad *wordnet*'ist küll süsteemse andmebaasi, kuid teisalt võib semantiliste suhete määramine keele mõistetele olla mõnevõrra ebasüsteemne tegevus. Nii võivadki minna omavahel vastuollu ühelt poolt ideaalpilt korrastatud *wordnet*'ist ja teiselt poolt tegelik töö – semantiliste suhete määramine. Selleks et kirjeldatud vastuolu ei tekiks või et tekkinud vastuolu parandada, kooskõlastavad *wordnet*'i tegijad semantiliste suhete määramise kriteeriume nii detailselt kui võimalik.

Siinne artikkel käsitlebki mõistete semantiliste suhete määramise ja kirjeldamise probleeme, millele Eesti Wordneti tegijad mõistelise andmebaasi koostamisel lahendusi otsivad. Kirjutise eesmärk on anda ülevaade Eesti Wordneti hetkeseisust, keskendudes selles kasutatavatele semantilistele suhetele. Kirjutises ei püüta ümber lükata *wordnet*'i koostamise eeskujude, Princetoni WordNeti ja EuroWordneti põhimõtteid, pigem mõtiskletakse selle üle, kas eesti keelt iseloomustavad eripärased suhted ja millised on need, mis peaksid tingimata ka Eesti Wordnetis kajastuma.

Selleks, et vähendada *wordnet*'i koostajate subjektiivset tõlgendust semantiliste suhete määramisel, soovitatakse nende suhete kontrollimiseks ja kindlakstegemiseks kasutada testlauseid (vt nt Hicks, Herold 2009). Semantiliste suhete testlausete rakendamine aitab kaasa *wordnet*'i ühtlase taseme saavutamisele, neid teste võib pidada üheks *wordnet*'i koostamise juhendi osaks. Eesti Wordneti koostajad peavad väga oluliseks tagada andmebaasis semantiliste suhete süsteemsus.

Artikli algusosas antakse ülevaade Eesti Wordnetist ja selle taustast. Edasises keskendutakse Eesti Wordnetis sisalduvatele põhilistele semantilistele suhetele. Esiteks käsitletakse üht osa semantilisi suhteid, mis esilduvad sama hierarhia sees – sünonüümiat, hüperonüümiat, meronüümiat. Seejärel käsitletakse neid semantilisi suhteid, mis esilduvad hierarhiate vahel ja näitavad mõiste(te) lisarolli või -funktsiooni – rollisuhet, hägussuhet. Artikli kokkuvõtteosas arutletakse selle üle, millised probleemid vajaksid Eesti Wordneti arendamisel veel lahendamist.

Varasemad Eesti Wordneti käsitlused keskenduvad üldiselt selle loomisloole (Vider jt 2000), andmebaasi ülesehitusele ja mahu suurendamisele (Orav jt 2011; Pedersen jt 2013). Siinne artikkel keskendub Eesti Wordneti semantilistele suhetele ning lähtekohaks on võetud fakt, et mõistetevaheliste suhete määramise probleemid saavad peaasjalikult alguse *wordnet*'i

põhiüksuste, sünohulkade tähenduslikust ebamäärasusest. Seega puutuvad *wordnet*'i koostajad sageli kokku nii mõistete tähenduste ebamäärasusega ja hägususega kui ka neid omavahel ühendavate leksikosemantiliste suhete hägususega. Ähmased piirid tähenduste vahel ja semantiliste suhete määramise keerukus iseloomustavadki loomuliku keele olemust. Ometi on oluline saavutada äärmine korrastatus, et seda andmebaasis esitada.

2. Eesti Wordnet keeleressursina

Eesti Wordnet on leksikosemantiline andmebaas, mida koostatakse üldjoontes inglise Princetoni WordNeti põhimõtteid järgides. Eesti Wordneti loomist alustati aastal 1995 Tartu Ülikoolis ning praeguseks sisaldab see üle 72 000 mõiste (sh sõnu u 98 700) ja üle 230 000 semantilise suhte. Sõnaliikidelt koosneb Eesti Wordnet adjektiividest, substantiividest, verbidest ja adverbidest, mis iga sõnaliigi sees on koondatud paljudesse tähenduslikesse üksustesse ehk sünohulkadesse (ingl *synset*) vaikimisi täis- ja lähisünonüümia suhte abil.

Siiani kulgenud tööperiood jaguneb laias laastus kaheks etapiks. Alustasime EuroWordNeti (EWN) projektis etteantud baasmõistete tõlkimisega (vt lähemalt Vider jt 2000; Orav jt 2011), mida laiendasime korpuse sagedusloendite järgi. Teine etapp algas 2007. aastal ja kestab Eesti riikliku keeletehnoloogiaprogrammi raames siiani ning enamik tesauruse suurusest ongi saavutatud tänu sellele programmile.

Wordnet'i kui väärtusliku keeleressursi tõusmine infotehnoloogia valdkonda on põhjustatud eelkõige vajadusest selgitada arvutisüsteemidele loomuliku keele mõisteseoseid, st arvuti peaks keeleandmete põhjal oskama ka teatud semantilisi järeldusi teha (nt mets koosneb puudest, pahtel on teatud ehitusmaterjal). *Wordnet*'ide eeliseks paljude teiste sõnastike ees on mitmekeelsus – eri keelte *wordnet*'id on omavahel ühendatud keeltevahelise indeksiga, mis võimaldab mõistepõhiselt tõlkevasteid leida. Kõige esimesele ehk ingliskeelsele *wordnet*'ile on viimaste aastate jooksul tulnud lisa üle kuuekümne keele kohta, sh näiteks ka surnud ladina keele *wordnet*[1].

Võrreldes Princetoni WordNetiga (vt Miller jt 1990) on Eesti Wordnetis märksa rohkem eri tüüpi semantilisi suhteid, et veelgi täpsemalt anda edasi

[1] Ladina keele *wordnet*'i kohta vt multiwordnet.fbk.eu.

tähendusnüansse. Kuna Eesti Wordnetis on olemas võimalus siduda ka eri sõnaliikidest lähtuvaid mõisteid, siis moodustub neist mõistetest mingi konkreetne semantiline väli, valdkond – semantiliselt seotud sõnade hulk, mis moodustab teatud mõistelise terviku (Õim 1997). Tänapäeva keeletehnoloogilised rakendused töötavad paremini paljuski just valdkondliku lähenemisega: näiteks sõnatähenduste ühestamine, masintõlge või infootsing saavad tänu tihedale semantilisele võrgustikule palju rohkem materjali, millega töötada (Vossen 1998). Samuti on *wordnet* oluline keeleteaduses näiteks keele leksikaalse struktuuri uurimisel, keeletehnoloogias tekstide automaatse kokkuvõtte tegemisel, sõnavaliku vigade automaatsel parandamisel tekstis jm (vt Budanitsky, Hirst 2006: 13).

Kõigi *wordnet*'ide eelkäijaks peetakse küll inglise Princetoni WordNeti, kuid nende koostamis- ja loomispõhimõtted pole alati samad. *Wordnet*'ide loomiseks on kasutatud eri strateegiaid – käsitsi, automaatset ning poolautomaatset loomist (Pedersen jt 2013). Keeled on valinud omakeelse *wordnet*'i ehitamiseks strateegiaid, mis ainult mõnel juhul toetuvad mingile teoreetilisele alusele. Kõige sagedasem meetod on olnud Princetoni WordNeti tõlkimine kas tervenisti või osaliselt. Nii on näiteks meie naabritel soomlastel oma *wordnet*'i koostamisel tõlgitud ära vaid sünohulgad, kuid nende vahele on taotluslikult jäetud samad semantilised suhted mis Princetoni WordNetis, samuti on mõistete seletused endiselt ingliskeelsed (Niemi & Linden 2012). Ka on kirjeldatud automaatsest loomisest tulenevaid probleeme näiteks horvaadi ja ühe portugali *wordnet*'i näitel, mis on loodud n-ö adapteeritud mudelina, võttes aluseks Princetoni WordNeti (vt Oliviera, Gomes 2014; Šojat, Srebačić 2014).

Eesti Wordnet on koostatud peamiselt käsitsi ja eesti keele eripära arvesse võttes. Tesaurust täiendatakse nii korpusloendite toel kui ka süvitsi mõne valdkonna sõnavaraga tegeldes. Kui võrrelda Eesti Wordneti tähendusi nende andmetega, mis sisalduvad ühestatud sõnatähendustega korpuses, võib järeldada, et üldkeeles enim esinev sõnavara ning sõnade tähendused on Eesti Wordnetis olemas. Seda ka sõnaliigiti – kõik sagedasemad adjektiivid ja adverbid oma paljude tähendustega on praeguseks lisatud. Eesti Wordnet on juba ületanud üldkeele sõnavara peegeldamise faasi ja sisaldab kohati väga spetsiifilisi erialaseid mõisteid. Paar katset on tehtud ka andmebaasi automaatseks suurendamiseks *mine*-tegevusnimede, *ja*-tegijanimedega jm (vt Orav jt 2011). See omakorda on tekitanud koostajaile rohkelt lisatööd, sest automaatselt lisatud tuletised on tarvis

käsitsi üle kontrollida ning semantiliste suhete kaudu omavahel seostada (lähemalt vt Kahusk jt 2010).

Kõigi *wordnet*'ide eelkäija, Princetoni WordNeti suurus – üle 120 000 mõiste – on paljudele keeltele veel kättesaamatu, sest enamiku keelte jaoks pole sobivaid arvutiressursse, mis võimaldaksid automaatselt tekitada mõisteid koos semantiliste suhete ja ingliskeelsete vastetega. Eri keelte *wordnet*'ide koostamisel keskendutakse mahu suurendamise kõrval samaväärselt kvaliteedi parandamisele. Ideaalis peaksid *wordnet*'is sisalduma leksikaalsed üksused, mis täidavad kolme eesmärki:

1) esinevad reaalsetes tekstides piisava sagedusega;
2) on olulised eri rakendustele (näiteks masintõlkes vajatakse adverbe kindlasti märksa rohkem kui infootsingus või refereerimises ning adverbide tähenduste eristamine ning nende õiged suhted aitavad eesmärki kergemini saavutada);
3) toetavad keeltevahelisi uuringuid ja rakendusi, mistõttu peaks olema tagatud *wordnet*'ide ühendamise võimalus (tavaliselt ingliskeelse Princetoni WordNeti kaudu). Suur osa eri keelte *wordnet*'idest on omavahel ka ühendatud inglise keele kaudu, st on võimalik teada saada, kas keeles esineb vastav mõiste, ning võrrelda selle mõiste seoseid teiste mõistetega. Näiteks eesti keeles on mõistel 'leib' 14 alammõistet (*peenleib*, *rukkileib*, *kuivikleib*, aga ka *armulaualeib*, *pruudileib* jne), kuid taanikeelsel vastel 'brød' ('leib') leiame lausa 34 alammõistet. Selline võrdlus annab hea võimaluse vaadata üle eestikeelsed mõisted ja vajaduse korral puuduvad lisada.

Princetoni WordNet on olnud vajalikuks ressursiks tuhandetele rakendustele. Tema populaarsus tuleneb nii mahust kui ka suhete süsteemist. Eesti Wordneti rakendamine seisab veel ees, kuigi järjest rohkem plaanitakse kasutada seda ressursina nii Eesti riikliku keeletehnoloogiaprogrammi projektides[2] kui ka keeletehnoloogilisi ressursse kasutavates ettevõtetes.

[2] Vt lähemalt Eesti riikliku keeletehnoloogiaprogrammi kodulehte www.keeletehnoloogia.ee.

3. Leksikosemantilised suhted Eesti Wordnetis

Juba alates Ferdinand de Saussure'i ideedest („Cours de linguistique générale", 1974/1916) on räägitud vajadusest kirjeldada sõnu paradigmaatiliste suhete kaudu – keelelisi üksusi saab määrata suhete kaudu teiste keeleliste üksustega sama süsteemi sees. Diskuteeritakse ka selle üle, millised suhted on leksikonis esmatähtsad, kuidas neid kindlaks teha ja kuidas tagada neist ühtmoodi arusaamine. Ehkki leksikosemantilistest suhetest on keeleteaduses palju kirjutatud (nt Lyons 1977; Cruse 1986, 2004; Murphy 2003; Budanitsky, Hirst 2006; Geeraerts 2010; Langemets 2010), on segadus püsinud.

Mõisteid kategoriseeritakse eri eesmärkidel. Leksikoloogias on kategoriseerimise eesmärk püüd saavutada süsteemsus, mis omakorda on aluseks näiteks sõnaraamatutele ja andmebaasidele. Mõistelise sõnaraamatu süsteem põhineb enamasti semantilistel suhetel, mis seovad sõnu ja mõisteid tähenduse alusel. Keele mõistete ja sõnade tähendustega tegelevad psühholoogid ja keeleteadlased on korduvalt tõdenud, et mõistete kategoriseerimine semantiliste suhete abil ei ole kerge ülesanne, sest esineb lahkarvamusi võimalike suhtetüüpide sisus ja arvus (Evens 1988). Oswald Werner (1988) väidab, et kõiki teadmisi on võimalik esitada ainult kolme suhte seisukohast: modifikatsiooni ehk teisenduse, taksonoomia ja järjestamise (ingl *sequencing*) järgi. Igor Melčuk ja Aleksandr Žolkovsky (1988) esitavad 53 suhtetüüpi, samal ajal kui Thomas Ahlswede ja Martha Evens (1988) kasutavad üle 100 suhte ainuüksi adjektiivide jaoks.

On arutletud, et suhtetüüpide arv sõltub paljuski just sõnaliigi eripärast. Adjektiividel võibki tähendus muutuda olenevalt sellest, millist substantiivi tähenduskomponenti see rõhutab ja millist tüüpi substantiivi täiendina esineb (Tuulik 2014: 307). Samuti kannab grammatiline vormistus alati mingit osa tähendusest (Pajusalu 2009: 82). Seega võib öelda, et olenemata leksikaalse semantika teooriate rohkusest on selge, et püüded semantiliste jaotuste vallas kukuvad mitmeski mõttes läbi, kui neid rakendatakse relatsioonilises võrgustikus ehk siis praktilisel tasandil. Siiski vajavad paljud keeletehnoloogia rakendused spetsiifilist semantilist esitust – sellist, mis võtaks arvesse mõiste koos tema kõigi võimalike semantiliste suhetega.

Leksikosemantilised seosed on keeleteadusest ja leksikograafiast tuttavad, seal rakendatakse neid leksikaalsete üksuste vahel. *Wordnet*'is on semantilised suhted pigem leksikaliseerunud mõistete vahel, ehkki

peaaegu kõigis *wordnet*'ides eristatakse ka antonüümiasuhet leksikaalsete üksuste vahelisena.

Kuigi *wordnet*'is kasutatavate suhete nimetused on samad kui leksikograafias üldiselt, pole lihtne kanda keeleline arusaam sõnade (või sõnaühendite) semantilisest seosest üle mõistetele või sünohulkadele. Olulisemad suhted *wordnet*'is, mida nimetatakse ka põhisuheteks (ingl *constitutive relations*), on sünonüümia, hüperonüümia ja osa-terviku suhted. Poola lingvistid on *wordnet*'i ülesehituse kohta piltlikult öelnud: „Sünonüümia moodustab *wordnet*'i nurgakivi, hüperonüümia tema selgroo ning osa-terviku suhe vajaliku ühendusliimi" (Maziarz jt 2013). Need suhted määravadki ära kogu *wordnet*'i konstruktsiooni iseärasuse.

Samuti on Eesti Wordnetis laias laastus kahte tüüpi semantilisi suhteid. Esimesed neist on põhisuhted, mille abil ühendatakse kaks sünohulka (nt hüperonüümia, osa-terviku suhe). Teine tüüp suhteid on leksikaalsed suhted, mille abil ühendatakse kaks sõna (nt antonüümia).

Semantiliste suhete täieliku nimekirja sai Eesti Wordnet kaasa EuroWordNeti projektist ja seda pole siiani muudetud, sest on paljuski tingitud *wordnet*'i töövahendi jäikusest, aga ka segadusest suhete olemusest arusaamisel. Kokku on Eesti Wordnetis 51 suhtetüüpi, sh põhitüüpide alltüübid. Alljärgnevas tabelis on toodud peamised suhtetüübid kõrvutatuna nii Princetoni WordNeti, EuroWordNeti kui ka Eesti Wordneti põhjal. Neist viimases kahes eristatakse meronüümiat ja holonüümiat ning rollisuhte puhul ka vastavaid all-liike (agenti, kohta, materjali jms).

Princetoni WordNet	**EuroWordNet**	**Eesti Wordnet**
sünonüümia	sünonüümia lähisünonüümia	sünonüümia lähisünonüümia
antonüümia	antonüümia lähiantonüümia	antonüümia lähiantonüümia
hüponüümia- hüperonüümia	hüponüümia- hüperonüümia	hüponüümia- hüperonüümia
meronüümia- holonüümia	meronüümia-holonüümia	meronüümia- holonüümia
troponüümia	verbide hüponüümia- hüperonüümia	verbide hüponüümia- hüperonüümia
põhjussuhe (*causes)*	põhjussuhe	põhjussuhe

tuletussuhe (*derived from*)	tuletussuhe	tuletussuhe
pertainymy	–	–
sarnane (*similar to*)	–	–
partitsiip	–	–
entailment	rollisuhe (*role*)	rollisuhe
–	*has_subevent*	*has_subevent*
atribuudisuhe	*be_in_state*	*be_in_state*
vaata ka	*fuzzynymy*	*fuzzynymy*

Princetoni WordNetis mõeldakse pertonüümide (ingl *pertainymy*) all nimisõnast tuletatud omadussõnu, nagu *finantsiline*, *intellektuaalne*. Iga pertonüüm on seotud nimisõna või tegusõnaga, mis EuroWordNetis ja Eesti Wordnetis on seostatud leksikaalse suhte 'derived from' ('tuletatud') kaudu, nt omadussõna *finantsiline* 'is_derived_from' ('on tuletatud') nimisõnast *finantsid*.

Princetoni WordNetis pole verbileksikoni hüperonüümia ka siiski päris noomenite oma sarnane. Seal nimetatakse vastavat seost troponüümiaks, mis näitab, mil viisil on üks verb spetsiifilisem ja tähenduse poolest kitsam kui teine (Beckwith jt 1990). Eesti Wordnetis, aga ka näiteks saksa GermaNetis (Kunze 1999) ja poola plWordNetis (Piasecki jt 2009) sellist eristust ei tehta, tegemist on üldiselt ikkagi hüperonüümiaga. Näiteks verbi *kõndima* alammõisteteks on *sumpama*, *tammuma*, *marssima*, *tuikuma*, *lonkima* jne.

Wordnet'i kasutamine mitmes loomulikku keelt töötlevas rakenduses on selgelt näidanud hierarhiliste suhete olulisust. Seda, millised teised suhted arvesse võtta, pole kerge otsustada, sest pole ühtset universaalset leksikograafilist kriteeriumi. On ka selge, et suhted on keeliti väga erinevad (Cruse 2004: 143).

Võrreldes Princetoni WordNetiga (Maziarz jt 2013 andmete põhjal) on Eesti Wordnetis kasutusel rohkem suhtetüüpe – kokku 51. See on ka põhjus, miks Eesti Wordnetis on hüperonüümia kasutusprotsent kõigest ligikaudu 28 võrreldes Princetoni WordNetiga, kus on kõige olulisem just hüperonüümia, mis moodustab 77,2% kõikidest suhetest. Princetoni WordNetis võib veel ära märkida osa-terviku suhte suurema osakaalu võrreldes Eesti Wordnetiga, muud suhted moodustavad Princetoni WordNetis

väikese osa. Eesti Wordnetis on aga ka kasutusel üpris palju rollisuhteid ning hägussuhteid, mida Princetoni WordNetis ei leidugi.

Semantilistel suhetel on nii teoreetiliselt kui ka praktiliselt keskne roll *wordnet*'i ülesehituses ja andmebaasi rakendustes. Eesti Wordneti koostamisel kasutatud leksikosemantilised suhted võimaldavad teostada mõistelise sõnaraamatu ideed – kõik keele mõisted on võimalik omavahel siduda semantiliste suhete abil. Mõnda neist suhetest on sünohulkade sidumisel hõlpsam määrata, mõnda mitte. Edasises keskendutakse suhetele, mis Eesti Wordneti koostamisel on kõige rohkem arusaamatusi tekitanud.

3.1. Sünonüümia

Sünonüümia on fundamentaalne suhe enamiku *wordnet*'ide jaoks. Princetoni WordNetis nimetatakse sünonüümiat „põhiliseks semantiliseks suhteks", sest sünonüümsed sõnad moodustavad sünohulga ja kõik selle liikmed osutavad leksikaliseerunud mõistele. Sünohulga moodustavad keeles iseseisvalt eksisteerivad täistähenduslikud sõnad, ainult et sünohulga piires peaksid nad olema mõisteliselt ekvivalentsed.

See, et sõnu võib sünonüümsuse alusel tervikmõisteteks koondada ja ülejäänud suhteid (nagu hüponüümia-hüperonüümia) viimaste najal käsitleda, luues niimoodi kirjeldatavas keeles realiseeritud mentaalsest leksikonist suhteliselt selge ja korrastatud pildi, oligi George Milleri alusidee (1998). Sünohulkade loomise ja sellest hargneva tähenduspõhise hierarhilise esituse idee levis kiirelt, kuigi mõistelisi, keelest sõltumatuid tesauruseid oli püütud teha varemgi (nt Roget' tesaurus[3]). Esialgu selge ja süsteemsena näiv sünohulkadest hargnev tähendusvõrgustik on ülejäänud suhete määramisel pigem probleemide allikaks. Üks neist probleemidest ongi hägusus, selgelt määratavate piiride puudumine.

Ühes sünohulgas olevate sõnade ehk täissünonüümide vahetuse puhul peab tähendus kontekstis jääma samaks (Miller 1998: 23; Cruse 2002: 489). Perfektset sünonüümiat esineb aga loomulikus keeles äärmiselt harva – stiili- ja tähendusvarjundeid on liialt palju – seetõttu sisaldab leksikon hulganisti sünonüüme, mis on piiratud asendatavusega. Enamikus *wordnet*'ides on kasutusel ka osa- või lähisünonüümia seos, mis leiab sagedast kasutust. Siia kuuluvad mitmesugused stiilivarjundid, intensiiv-

[3] Roget's International Thesaurus of English Words and Phrases. New York: Thomas Y. Crowell, 1922.

susastme erinevus (*ilus – kena*), sotsiaalsete gruppide erinevused (*ema – mutt*) jms. Osa- või lähisünonüümia (Princetoni WordNetis on suhte nimetus 'see_also'; EuroWordNetis ja Eesti Wordnetis 'near_synonym') tähendab, et sünonüümsete sõnade tähendused langevad kokku vaid mõnes kontekstis või seal, kus nende omavaheline asendamine ei muuda lause tõeväärtust, nt *kosmoselaev – kosmoserakett*, *nässu – katki*, *küna – lootsik*, *ajakirjanik – saatejuht*. Aga näiteks sünonüümipaaris *nali – temp* tuleb sõna *temp* tähenduses esile tahtlik planeeritud tegevus, kuid nali iseenesest võib olla ka juhuslik.

Semantilise sarnasuse kindlakstegemiseks kasutatakse mitmesuguseid meetodeid. Saab rakendada lingvistilisi kontrollteste, nagu on pakkunud lingvistid (nt Cruse 1986): *kui ta/see on X, siis ta/see on ka Y*. On võimalik teha teste keelekasutajaid intervjueerides, kuid see töö on äärmiselt ressursirohke. Küsitlustele lisaks on tulnud järjest rohkem arvutiteaduslikke lähenemisi, nagu distributiivne semantika kui leksikaalse semantika tööriist, mis võimaldab konstrueerida tähenduskirjeldusi sõnade esinemuse põhjal ulatuslikes tekstikorpustes. Rakenduslikust küljest saab esile tuua näiteks töövahendi WordNet: Similarity (Potsma, Vossen 2014), millega saab mõõta semantiliselt sarnaste mõistete kaalu. Viimased on keelest sõltumatud mõõtmisviisid, kuid väidetavalt toetavad paljuski inimese keelelist intuitsiooni.

Võrdselt keeruline on defineerida sünohulka sünonüümia kaudu ja sünonüümiat sünohulga kaudu (Maziarz jt 2013). Artiklis „On wordnets and relations" (Piasecki jt 2013) kirjutatakse, et sünohulga moodustamine on osutunud raskeks ülesandeks ja seetõttu loodetakse pigem *wordnet*'i koostaja intuitsioonile. Intuitsioon on aga teadupärast kõigil erinev ja seetõttu on ka sünohulki moodustatud üsna erinevalt. Selle tõttu on *wordnet*'i koostajad nõrgendanud sünonüümia määramise kriteeriume. Sünohulga definitsioon on siinses kontekstis järgmine: kogum (lähi)sünonüüme, mis osutab leksikaliseerunud mõistele ja mille tähendust jagavad kõik sünohulga liikmed. Lisakriteeriumina peavad sünohulgad jagama samu hüponüüme ja hüperonüüme ning holonüüme ja meronüüme (Piasecki jt 2009). Seega on sünohulgad moodustatud nii täissünonüümia kui ka osasünonüümia alusel.

Sünohulkade tekitamisel on oluline meeles pidada ka kriitikat *wordnet*'ide liigse granuleerituse ehk üleeristamise kohta (Jiamjitvanich, Yatskevich 2009). Kui kõik stiilivarjundiga ning peaaegu sarnase või

osasünonüümiaga tähendused eri mõistetesse liigitada, siis muutuvad tähendused liialt üleeristatuks ning *wordnet* loomuliku keele rakendustes keeruliselt kasutatavaks. Näiteks SentiWordnet[4], millel on infot ka emotsionaalse hinnangu kohta (st mõistel on juures märgend positiivne, negatiivne või neutraalne) ja mida kasutatakse tekstide meelestatuse analüüsis, jääb kimpu emotsionaalsuse määramisega, sest ei suuda tähendusi eristada.

3.2. Hüponüümia ja hüperonüümia

Kirjeldada sõnu või mõisteid ülemmõistete kaudu on sõnastikes üsna tavapärane moodus, kus seletustes on peasõnaks tihti hüperonüüm, ehkki võib leiduda ka hüponüüme (Svensén 2009: 218–219, 249). Tuntud leksikograaf Sue Atkins on osutanud, et oleks ideaalne, kui kõik seletused oleksid antud ülemmõistega (Atkins, Rundell 2008: 146). Ka *wordnet*'is on peale sünonüümia hierarhilised suhted kõige olulisemad ja seda mitmel põhjusel. Üks põhjusi on inimeste vajadus struktureeritud info järele. Teine põhjus on rakenduslik – arvutiprogrammid vajavad infot just maailmateadmiste liigitamise kohta, et saaks hõlpsasti liikuda üldisemalt spetsiifilisemale. Näiteks kui infootsiprogrammiga otsida sõna *mamba*, saame tulemuseks, et tegemist on roomajaga ja see kuulub ühte klassi teiste roomajatega, nagu *püüton*, *varaan*. Kolmandaks on arvutisõnastikus oluline andmete töödeldavus ehk siis ei piisa ainult definitsioonis olevast ülemmõistest, mida arvuti ei suuda kiirelt üles leida.

Hüponüümia ja hüperonüümia on semantiline suhe, mis esildub nimisõnade, tegusõnade ja osaliselt ka omadussõnade vahel. Selle suhte kindlakstegemiseks kasutatakse järgmisi testlauseid.

> X on Y, aga Y pole ainult X.
>
> *Kask on puu, aga puu pole ainult kask.*
>
> Kui A pole Y, siis ei saa ta olla ka X.
>
> *Kui kass pole loom, siis ei saa ta olla ka lemmikloom.*

Uurimustes (nt Hicks ja Herold 2009) on osutatud sellele, et mõni hüperonüüm on oma olemuselt jäigem (ingl *rigidity*) ja mõni mitte. Kindlakstegemiseks kasutatakse testküsimusi.

---

[4] SentiWordneti koduleht, vt http://sentiwordnet.isti.cnr.it/.

Kas X on alati või ilmtingimata Y?

Kas Xi saab peatada olemast Y?

Nii on kass alati teatud loom ja mõnel juhul võib ta olla ka lemmikloom (kuigi mitte alati, sest nt hulkuvad kassid ei ole mitte kellegi lemmikloomad). Sellisel juhul võib kass olla lemmiklooma rollis, st on hoopis seotud rollisuhtega.

Mõisteid, mis paiknevad sama taseme hüponüümidena ühe ja sama hüperonüümi all, nimetatakse kaashüponüümideks (ingl *co-hyponym*). Näiteks *harakas*, *ronk* ja *varblane* on omavahel kaashüponüümid ja paiknevad ühise hüperonüümi *lind* all. Kaashüponüümia rusikareegel on seega *X ja Y on mõlemad Z-id*. Kuidas eristada omavahel kaashüponüüme, peaks välja tulema teistest mõiste sisu edasiandvatest suhetest, kuid alati pole see nii. Näiteks on üldkeelse mõiste 'koer' alla koondatud *juhtkoer*, *ajukoer*, *jahikoer*, *õuekoer*, *toakoer* – teatud funktsiooniga koeratüübid, ja neile lisaks *koeralita*, *spits*, *krants*. Ehkki sedasorti kaashüponüümid ei jaga alati samu selgeid kriteeriume, et ühes taksonoomias esineda, on Eesti Wordnetis need siiski koos. Hägususest hoolimata ei ole neid võimalik ka mujale hierarhiasse panna, kuna kuuluvad samasse semantilisse välja.

Mõnikord pole hierarhilise kooskõla jaoks keeles olemas leksikaliseerunud mõisteid, mistõttu võib loogilisse järjestusse jääda tühimik. Võimaluse korral täidetakse see tehislikult loodud vahemõistega. Näiteks on Eesti Wordnetis paljude alammõistetega sünohulgad 'seisund, seisukord, seis, olek', mis moodustab hierarhiapuu 102 esimese astme alammõistega (kõiki alammõisteid kokku on 2761). Sünohulk 'inimene, inimolend, indiviid, isik' moodustab hierarhiapuu 431 esimese astme alammõistega (kokku 6424 alammõistet). Selleks, et koondada hierarhiasse tähendusvälja poolest kokku kuuluvaid sõnu, loodi neis hierarhiais vaheastmed 'psüühiline seisund' (mille all on *paanika*, *katarsis*, *amneesia*, *enesekindlus* jms) ning 'õnnetu inimene' (koos sõnadega *hädavares*, *hädasolija*, *nohik*, *patuoinas* jms).

Teinekord võib tekkida probleeme ka hüperonüümia määramisega, täpsemalt sobivaima ülemmõiste valimisega, sest püütakse jälgida, et võimaluse korral oleks igal mõistel vaid üks hüperonüüm (Atkins, Rundell 2008: 146). Näiteks mõiste 'neuropediaatria' puhul tekib küsimus, kas selle arstiteaduse haru ülemmõiste oleks pigem *pediaatria* või *neuroloogia*. Eeldatakse, et liitsõna teine osis on ülemmõiste, aga „Meditsiinisõnas-

tiku“ seletuses väidetakse, et tegu on pigem neuroloogia haruga, millega pediaatrid tegelevad.

Samamoodi kui saksa GermaNetis rakendatakse Eesti Wordnetis hierarhilist seost adjektiivide puhul. Kui GermaNetis on hierarhiasse pandud kõik adjektiivid, siis eesti omas ainult üksikud, tavaliselt liitsõna ühe osise järgi (nt *kadeda* alammõiste on *armukade*; *lahke* alammõiste *külalislahke*). Samasugust loogikat on järgitud ka adverbide puhul, nt *kadedalt* ja *armukadedalt*.

3.3. Holonüümia ja meronüümia

Meronüümia ja holonüümia on osa-terviku suhe ning koos hüponüümia-hüperonüümiaga koondab see mõisted küll ühisesse semantilisse välja, kuid pole leksikograafias nii sagedasti kasutusel kui sünonüümia ja muud hierarhilised suhted (Murphy 2003: 123).

Holonüümia ja meronüümia puhul on tegemist samuti hierarhilise leksikaalse suhtega, mille puhul iga alumise sõlme mõiste on enda suhtes ülemise mõiste osa. Alumine mõiste on endast vahetult ülalpool oleva mõiste suhtes meronüüm, ülemine mõiste alumise suhtes aga holonüüm. Loogiliselt järjekindla meronüümia puhul kehtivad järgmised testlaused.

Y on X-i osa;

X-il on Y;

X koosneb Y-i(de)st;

X-il on Y(-id).

Klassikaline näide on mõiste ’keha’ kui holonüüm ja kehaosi tähistavad mõisted kui selle meronüümid. Meronüümiale on iseloomulik pööratavus – kui *rool* on *auto* osa, siis *auto* osa on *rool*. Meronüümia puhul eristatakse mitmeid alltüüpe, nagu näiteks: ’on millegi osa’ (*kõvaketas – arvuti*), ’on liige’ (*sõdur – sõjavägi*), ’on tehtud materjalist’ (*puu – uks*), ’on mingi koha osa’ (*klass – kool*) ja ’on osa mingist portsjonist’ (*leivakäär – leib*).

Mõnikord on raske teha vahet hüperonüümia ja meronüümia vahel, näiteks seltsi, liiki kuuluvate taimede, lindude jt puhul. Tekib küsimus, kas *flamingo* on *flamingolaste* hüponüüm või on üks osa *veelindudest*? Või siis on *flamingo* pigem *veelinnu* hüponüüm ja üks *flamingolaste*

hulka kuuluv linnuliik? Princetoni WordNetis on näiteks konkreetsed linnuliigid *liigi* hüponüümid ning üldkeelse mõistega 'lind' on nad seotud üpriski ebaühtlaselt. Ka Eesti Wordnetis on olukord nii lindude, loomade kui ka taimedega ebaselge. Probleem seisneb osalt just selles, et Eesti Wordnet on üldkeelest arenenud palju kaugemale ja spetsiifilisematesse valdkondadesse. Siiski tuleb ka spetsiifilised mõisted siduda üldkeele mõistetega.

3.4. Rollisuhe

Kasutussageduselt teisel kohal olev suhe Eesti Wordnetis on rollisuhe. Princetoni WordNetist rollisuhet ei leia, see võeti eestikeelse andmebaasi jaoks kasutusele EuroWordNeti projekti raames. Rollisuhe on oluline funktsionaalne suhe, mille abil on võimalik mõiste sisu veelgi täpsemini avada. Semantiline rollisuhe on suhe nimisõna ja tegusõna vahel, aga ka nimisõna ja nimisõna vahel. Tihti ei pruugi pelgalt hüperonüümia olla piisavalt informatiivne, näiteks mõisted 'ristima' ja 'vader'. Mõiste 'ristivanem' puhul iseloomustab sellele määratud rollisuhe (agendiks on *vader*, *ristivanem*) verbi *ristima* rohkem kui viimase ülemmõisteks olev *panema* (tähenduses 'määrama kedagi kellekski'). Rollisuhet määratakse olenevalt situatsioonist, võimalikud suhted on järgmised.

- Agent – temaatiline roll, mis osutab tegevuste, protsesside ja seisundite agentidele. Eesti Wordnetis on *ja*-tegijanimed verbidest automaatselt moodustatud, mistõttu on suhe tegijanime ja verbi vahel enamasti süstemaatiline (*liikuma – liikuja*).
- Patsient – keegi/miski, kes midagi läbi teeb või läbi elab (*õppima – õppija*).
- Koht – kus midagi juhtub (*õpetama – kool*).
- Instrument – osutab vahendile, millega mingit tegevust ellu viiakse (*haamer – haamerdama*; *naelutama*).
- Suund – koht, kust või kuhu liikumine on suunatud (*külastama – koht*).
- Viis – kuidas midagi toimub (*norskab – kõvasti*).
- Lähtekoht – kust midagi saab alguse (*võistlema – start*).
- Sihtkoht – koht, kuhu keegi satub või midagi teeb (*põrand – kukkuma*).

Neid all-liike võib olla rohkemgi, nt plWordNetis on situatsioonis osalevateks seosteks määratud veel lisaks produkt, aeg ja objekt (Maziarz 2011). Rollisuhete iseloomulikuks omaduseks on pööratavus. Näiteks kui mõistega 'kõnelema' kaasneb tegija *kõneleja*, siis kehtib pöördsuhe – *kõneleja* mängib rolli mõistes 'kõnelema'. Agendi ja instrumendi rollisuhte määramisel on Eesti Wordneti tegijad olnud üpris süstemaatilised. Küll aga on ebaühtlane kaasneva koha määramine – kaasneva kohana on *tootmise* puhul *tootmisruum* ja näiteks *pimesoolepõletiku* puhul *pimesool*, *seinataldriku* puhul on kaasnev koht *sein*. Piek Vossen (2002) pakub testi, kuidas kontrollida kaasneva koha suhte kehtivust: *X on koht, kus Y juhtub/toimub*. Seda järgides on võimalik välja arvata kaasneva koha suhtest viimane näide, kuid ilmselt on vaja leida sellele mõni muu sobivam suhe.

3.5. Hägussuhe

Assotsiatsioonid keele semantilises ruumis tunduvad olema lõputud. Hägussuhe (ingl *fuzzynymy)*, nagu nimetuski ütleb, on täpsustamata suhe ja osutab semantilisele assotsiatsioonile, mis on keelekõneleja jaoks ilmne, sest see võib vihjata teatud semantilisele väljale, millega mõiste seotud on (Svensén 2009: 210). Hägussuhet saab määrata ka sõnaliikide vahel, sellega võib ühendada näiteks nimisõna ja nimisõna, nimisõna ja omadussõna või nimisõna ja tegusõna.

Hägussuhteid esineb Eesti Wordnetis olevate mõistete vahel palju (u 9% kõigist määratud suhetest). Oleme arvamusel, et pigem fikseerida mõistete seotus hägussuhtena, kui jätta mõisted sidumata, sest neid suhteid on hiljem võimalik automaatselt muuta või eemaldada. Samas puuduolevate suhete lisamine kõigi andmebaasis olevate mõistete vahele on üsna ajamahukas töö. Kui hägussuhteid lähemalt uurida, selgub, et joonistuvad välja kindlad grupid, teatud tüüpi võimalikud suhted. Järgnevalt mõni näide selliste juhtude kohta.

- Suhted, mis on töötajate ja nende töökohtade vahel (*linnapea – raekoda*; *kuraator – muuseum*; *arst – haigla*; *kiirabiarst – kiirabi*; *sotsiaalpedagoog – kool* jms). Mõningatel juhtudel, kui mõiste looja arusaam ja keeleline vaist on nii öelnud, on neid käsitletud ka kui koha meronüümiat tähistavat suhet, nt *haigla* 'has_meronym_location' *arst*, kuid osa-terviku suhte määramise testlausete *Y on X-i osa*; *X-il on Y*; *Y koosneb X-i(de)st*; *Y-il on X(-id)* järgi *arst on*

haigla osa; *haiglal on arst*; *haigla koosneb arstidest*; *haiglal on arstid*, ning see ei tundu olevat korrektne suhe.

- Mõistepaar 'aednik' – 'aed' kuulub samuti siia kategooriasse, aga kas ka *aednik – aednikumaja*? Aednikumaja on maja, kus juhtumisi elab aednik ja ei pruugi kuuluda tingimata aedniku elukutse juurde.
- Suhted tegevuste ja nende toimumiskoha vahel (*spordiväljak – sport*; *promenaad – jalutuskäik*). Enne leksikosemantilise suhte määramist tuleb maailmateadmusele toetudes vastata küsimusele, kas sport on spordiväljaku osa või spordiväljakul on sport.
- Ainevaldkonnad, mis on seotud selle ala spetsialistidega või kohaga, kus sellega tegeletakse (*muuseum – museoloogia*; *bioloogia – bioloog* jms).

– Esemed, mis kuuluvad kellegi või millegi juurde ja kirjeldavad tema teatud eripärasusi (*postiljon – postikott*; *arst – stetoskoop*; *arvuti – arvutioskus*).
– Esemed, mille abil saab midagi parandada, muuta või teha (*kätgut* kui *haavaõmblusniit – haav*; *ilmutusaine*, *ilmuti – fotograafia*, *fotondus*).
– Ese mingiks otstarbeks, mingi funktsiooniga (*soova*[5] *– õllenõu*; *ravim – ravimiuuring*).
– Liigitamatud, kuid samasse tähendusvälja kuuluvad hägussuhtega mõisted (*kool –koolivorm*; *kool – koolitarbed*; *labor – laborihiir*).

Hägussuhe kipub esilduma eri sõnaliikide vahel ja eri semantiliste kategooriate vahel. Osa siintoodud hägussuhetest võiks muuta kas rollisuhteks või osa-terviku suhteks, nt töökoht ja töötaja, tegevus ja tegevuskoht, tegevus ja selle juurde kuuluvad atribuudid. Samas on Eesti Wordnetis mingis osas hägussuhetena määratud ka väga vabu tähendusseoseid (*arst – varesejalg*, *kirbukiri*), mille assotsiatsioon jääb kaugeks, mõisted ei kuulu samasse tähendusvälja ja seetõttu tuleks neid käsitleda kui vigu. Sellega seonduv töö kvaliteedi parandamise eesmärgil käib Eesti Wordneti täiendamisel pidevalt.

[5] Soova – õllenõu rest (EKSS).

4. Kokkuvõte ja edasised plaanid

Wordnet-tüüpi arvutisõnastikes on kõik mõisted omavahel ühendatud semantiliste suhetega. Eesti Wordneti maht on aasta-aastalt suurenenud, ületades praegu 72 000 mõiste piiri, semantilisi suhteid on nende vahel üle 230 000. Töö käigus on esile kerkinud vajadus semantiliste suhete määramist täpsustada, et saaks üle kontrollida andmebaasis kajastuv hetkeseis ja et juhised oleks uute mõistete sisestamiseks piisavalt selged. Artiklis käsitletud temaatika hõlmabki Eesti Wordneti hetkeseisu, mille põhjal on keskendutud semantiliste suhete määramise probleemidele. Eesti Wordnetti on algusest peale koostatud n-ö alt-üles-põhimõttel. See tähendab, et kõrvale on jäetud täisautomaatne koostamine – kõigile *wordnet*'idele eekujuks oleva Princetoni WordNeti tõlkimine. Pigem on Eesti Wordneti koostamisel lähtutud eesti keelele eripärastest semantilistest suhetest ning põhimõttest lisada mõistete vahele võimalikult palju semantilisi suhteid.

Artiklis leidsid käsitlemist leksikosemantiliste suhete erijuhud, mis mõnel juhul viitavad mõiste hägusale sisule ning mõnel juhul ka suhte enda mitmetimõistetavusele. Esitatud näidetega probleemid on aluseks Eesti Wordnetis sisalduvate vigade süstematiseerimisele ja parandamisele. Leksikosemantilised suhted erinevad küll keeliti, aga ka ühe keele *wordnet*'i koostajate endi arusaamad võivad suuresti erineda. Jõudsime tõdemusele, et oleks hea, kui leksikograafil oleks olemas juhendid, n-ö väljatöötatud lingvistilised testid, mis aitaksid mõistetevahelisi seoseid kinnistada. Ka siinse artikli tarbeks näidete läbivaatamine aitas andmebaasis olevat süstematiseerida ja üldist arusaama ühtlustada.

Suurimad probleemid semantiliste suhetega Eesti Wordnetis on olnud järgmised.

- Entsüklopeediline klassifikatsioon *vs.* üldkeel ja sellest eristusest tulenevad suhted.
- Sõnatähenduste vabu assotsiatsioone kiputakse käsitlema kui semantilist suhet ja need fikseeritakse kui hägussuhted, kuigi semantilise suhte lisamine taolisse andmebaasi on otstarbekas ainult siis, kui seostatavad mõisted kuuluvad samasse semantilisse välja. Samas on argument seegi, et suhteid andmebaasist eemaldada on lihtsam, kui neid juurde tekitada.
- Valdkonnasuhte puudumine Eesti Wordnetis. Valdkonnasuhe kannab sama ideed, mida leksikograafias kannavad üldiselt semantilised tüübid. Arvutileksikoloogilist eesmärki silmas pidades

võivad semantilised tüübid olla uurijale abiks polüseemsete sõnade analüüsil, eriti süstemaatilise polüseemia selgitamisel (Langemets 2010: 252). Ühtlasi võivad semantilised tüübid sõnaraamatutöös olla ka praktiliseks abivahendiks sõna semantika kodeerimisel. Kui teatud sõnade tähenduste vahelduses ilmnev regulaarsus on juba kindlaks tehtud, siis võib vastavaid malle kodeerides sõnaraamatus semantilise info esitust korrastada ning näidata ja selgitada sõnatähendustevahelisi loogilisi seoseid (Langemets 2010: 252; Tuulik 2014).

Eesti Wordneti koostamisel tuleks tulevikus silmas pidada kahte suuremat eesmärki – kvaliteetne mõisteline sõnastik eesti keele kohta ning kvaliteetne mitmekeelne arvutiressurss. Praegu keskendutaksegi eesti andmebaasi töös ingliskeelsete seoste kontrollile, sest üha suureneb huvi siduda omavahel eri keelte *wordnet*'e. Princetoni WordNeti suurus, 120 000 mõistet, on väga üksikute keelte puhul saavutatud, Eesti Wordnet oma mõistehulgaga on üsna keskmisel tasemel.

Töö Eesti Wordneti koostamisel jätkub nii sisuliselt (leksikosemantiliste suhete teooria ja praktikaga) kui ka mahuliselt. Loomuliku keele leksikosemantilist süsteemi pole mõttekas kirjeldada ainult analüütilisel teel, vaid pigem tuleks töötada pidevalt läbi kõik üksikjuhtumid praktilisel tasandil. Samuti tuleb *wordnet*'i headust pidevalt tõestada ja hinnata keeletehnoloogiliste rakenduste najal.

Kirjandus

Ahlswede, Thomas, Martha W. Evens 1988. A lexicon for a medical expert System. – Relational Models of the Lexicon. Ed. Martha W. Evens. New York: Cambridge University Press, 97–111.

Atkins, Sue, Michael Rundell 2008. Oxford Guide to Practical Lexicography. Oxford: Oxford University Press.

Beckwith jt 1990 = Richard Beckwith, Christiane Fellbaum, Derek Gross, George A. Miller. WordNet. A lexical database organized on psycholinguistic principles. – Using On-line Resources to Build a Lexicon. Ed. Uri Zernik. Hillsdale, NJ: Erlbaum, 211–231.

Bejar jt 1991 = Isaac I. Bejar, Roger Chaffin, Susan Embretson. Cognitive and Psychometric Analysis of Analogical Problem Solving. New York: Springer-Verlag.

Budanitsky, Alexander, Graeme Hirst 2006. Evaluating WordNet-based measures of lexical semantic relatedness. – Computational Linguistics 32 (1), 13–47.

Cruse, Alan D. 1986. Lexical Semantics. Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

Cruse, Alan D. 2002. Lexicology. An International Handbook On the Nature and Structure of Words and Vocabularies. 1. Walter de Gruyter GmbH.

Cruse, Alan D. 2004. Meaning in Language. An Introduction to Semantics and Pragmatics. New York: Oxford University Press.

EKSS = Eesti keele seletav sõnaraamat. 2009. „Eesti kirjakeele seletussõnaraamatu“ 2., täiendatud ja parandatud trükk. Toim. Margit Langemets, Mai Tiits, Tiia Valdre, Leidi Veskis, Ülle Viks, Piret Voll. Eesti Keele Instituut. Tallinn: Eesti Keele Sihtasutus. http://www.eki.ee/dict/ekss/ (30.04.2015).

Evens, Martha W. (ed.) 1988. Relational Models of the Lexicon. New York: Cambridge University Press.

Fellbaum, Christiane 1998. WordNet. An Electronic Lexical Database. Cambridge, MA: MIT Press.

Geeraerts, Dirk 2010. Theories of Lexical Semantics. Oxford: Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780198700302.001.0001.

Hicks, Amanda, Axel Herold 2009. Evaluating Ontologies with Rudify. – Proceedings of the International Conference on Knowledge Engineering and Ontology Development (KEOD ’09), Funchal – Madeira, Portugal, October 6–8, 2009. INSTICC Press, 5–12.

Jiamjitvanich, Kanjana, Mikalai Yatskevich 2009. Reducing polysemy in WordNet. – Proceedings of OM.

Kahusk jt 2010 = Neeme Kahusk, Kadri Kerner, Kadri Vider. Enriching Estonian WordNet with derivations and semantic relations. – Baltic HLT Proceedings: Human Language Technologies – the Baltic Perspective. Riga (Latvia) October 7–8, 2010. IOS Press (Frontiers in Artificial Intelligence and Applications), 195–200.

Kilgarriff, Adam 2000. WordNet. An electronic lexical database. Review. – Language 76 (3), 706–708. http://dx.doi.org/10.2307/417141.

Kunze, Claudia 1999. Semantics of verbs within GermaNet and EuroWordNet. – Proceedings of the workshop at 11th European summer school in logic, language and information. Ed. E. Kordoni, 189–200.

Langemets, Margit 2010. Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus keelevaras. Tallinn: Eesti Keele Sihtasutus.

Lyons, John 1977. Semantics. 1–2. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9781139165693, http://dx.doi.org/10.1017/CBO9780511620614.

Maziarz, Marek 2011. Semantic Relations among Nouns in Polish WordNet Grounded in Lexicographic and Semantic Tradition. (= Cognitive Studies 11.) http://www.site.uottawa.ca/~szpak/selected_publications_for_download/Wordnet/CS%2011%2010-Maziarz-Piasecki-Szpakowicz.pdf (29.12.2014).

Maziarz jt 2013 = Marek Maziarz, Maciej Piasecki, Stanisław Szpakowicz. The chicken-and-egg probleem in wordnet design: synonymy, synsets and constitutive relations. – Language Resources and Evaluation 47 (3), 769–796.

Meditsiinisõnastik = Meditsiinisõnastik. Eestikeelsed terminid koos seletuste ning ladina, inglise ja soome vastetega. 2004. 2., uuendatud trükk. Toim. Sirje Ootsing, Laine Trapido. Tallinn: Medicina.

Melčuk, Igor, Aleksandr Žolkovsky 1988. The explanatory combinatorial dictionary. – Relational Models of the Lexicon. Ed. Martha W. Evens. Cambridge: Cambridge University Press, 41–74.

Miller jt 1990 = George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine J. Miller. Introduction to WordNet. An on-line lexical database. – International Journal of Lexicography 3, 235–312.

Miller, George A. 1998. Nouns in WordNet. – WordNet. An Electronic Lexical Database. Ed. Christiane Fellbaum. Cambridge, MA: The MIT Press, 23–46.

Murphy, Lynne M. 2003. Semantic Relations and the Lexicon. Cambridge: Cambridge University Press. http://dx.doi.org/10.1017/CBO9780511486494.

Niemi, Jyrki, Krister Linden 2012. Representing the translation relation in a bilingual Wordnet. – Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC ’12), Istanbul, Turkey, 21–27 May 2012, 2439–2446.

Oliveira, Hugo Gonçalo, Paulo Gomes 2014. Onto.PT: recent developments of a large public domain Portuguese wordnet. Anthology. – Proceedings of the Seventh Global WordNet Conference (GWC 2014), Tartu, Estonia, January 25–29, 2014. Esd. Heili Orav, Christiane Fellbaum, Piek Vossen. Tartu: Tartu University Press, 16–22.

Orav jt 2011 = Heili Orav, Kadri Kerner, Sirli Parm. Eesti Wordneti hetkeseisust. – Keel ja Kirjandus 2, 96–106.

Pajusalu, Renate 2009. Sõna ja tähendus. Tallinn: Eesti Keele Sihtasutus.

Palmer, Martha 2009. Semlink. Linking PropBank, VerbNet and FrameNet. – Fifth International Workshop on Generative Approaches to the Lexicon (GL 2009). Pisa, Italy, 9–15.

Pedersen jt 2013 = Bolette S. Pedersen, Lars Borin, Markus Forsberg, Neeme Kahusk, Krister Lindén, Jyrki Niemi, Niklas Nisbeth, Lars Nygaard, Heili Orav, Hirkur Rögnvaldsson, Mitchel Seaton, Kadri Vider, Kaarlo

Voionmaa. Nordic and Baltic wordnets aligned and compared through „WordTies“. – Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013), May 22–24, 2013, Oslo, Norway. Eds. Stephan Oepen, Kristin Hagen, Janne Bondi Johannessen. (= NEALT Proceedings Series 16, Linköping Electronic Conference Proceedings 85.) Linköping: Linköping University Electronic Press, 147–162.

Piasecki jt 2009 = Maciej Piasecki, Stanisław Szpakowicz, Bartosz Broda. A wordnet from the ground up. Oficyna Wydawnicza Politechniki Wrocławskiej, Wrocław. http://www.plwordnet.pwr.wroc.pl/main/content/files/publications/A_Wordnet_from_the_Ground_Up.pdf (29.12.2014).

Piasecki jt 2013 = Maciej Piasecki, Stanisław Szpakowicz, Christiane Fellbaum, Bolette S. Pedersen. On wordnets and relations. – Language Resources and Evaluation 47 (3), 757–767.

Potsma, Marten, Piek Vossen 2014. What implementation and translation teach us. The case of semantic similarity measures in wordnets. – Proceedings of the Seventh Global WordNet Conference (GWC 2014), Tartu, Estonia, January 25–29, 2014. Eds. Heili Orav, Christiane Fellbaum, Piek Vossen. Tartu: Tartu University Press, 133–142.

Saussure, Ferdinand de 1974 (1916). Cours de linguistique générale. Payot, Lausanne, Paris.

Svensén, Bo 2009. A Handbook of Lexicography. The Theory and Practice of Dictionary-Making. Cambridge: Cambridge University Press.

Šojat, Krešimir, Matea Srebačić 2014. Morphosemantic relations between verbs in Croatian WordNet. – Proceedings of the Seventh Global WordNet Conference (GWC 2014), Tartu, Estonia, January 25–29, 2014. Eds. Heili Orav, Christiane Fellbaum, Piek Vossen. Tartu: Tartu University Press, 262–267.

Tuulik, Maria 2014. Adjektiivide polüseemia korpuses ja sõnaraamatus. – Eesti Rakenduslingvistika Ühingu aastaraamat 10. Toim. Helle Metslang, Margit Langemets, Maria-Maren Sepper. Tallinn: Eesti Rakenduslingvistika Ühing, 307–317. http://dx.doi.org/10.5128/ERYa10.19.

Vider jt 2000 = Kadri Vider, Neeme Kahusk, Heili Orav, Haldur Õim, Leho Paldre. Eesti keele tesaurus. – Arvutuslingvistikalt inimesele. Toim. Tiit Hennoste. (= Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1.) Tartu: Tartu Ülikooli Kirjastus, 127–152.

Vossen, Piek (ed.) 1998. EuroWordNet. A multilingual database with lexical semantic networks. Kluwer Academic Publishers Norwell.

Vossen, Piek 2002. EuroWordNet General Document. Version 3. Final. July 1, 2002. http://www.vossen.info/docs/2002/EWNGeneral.pdf (15.03.2015).

Werner, Oswald 1988. How to teach a Network. Minimal design Features for a cultural acquisition device or C-KAD. – Relational Models of the Lexicon. Ed. Martha W. Evens. New York: Cambridge University Press, 147–166.

Õim, Haldur 1997. Eesti keele mentaalse maailmapildi allikaid ja piirjooni. – Pühendusteos Huno Rätsepale 28.12.1997. Toim. Mati Erelt, Meeli Sedrik, Ellen Uuspõld. (= Tartu Ülikooli eesti keele õppetooli toimetised 7.) Tartu: Tartu Ülikooli Kirjastus, 255–268.

Fuzzy lexical-semantic relations in Estonian Wordnet

HEILI ORAV, SIRLI ZUPPING,
KADRI VARE

This paper gives an overview of the principles of wordnets in general and focuses mainly on the Estonian Wordnet (EstWN). The latest version of EstWN consists of more than 72,000 concepts and 51 different lexical relations are used to form a network of more than 230,000 semantic relations between concepts.

The main relations that EstWN uses are hyperonymy, meronymy, involvement and fuzzynyms (in Princeton WordNet, for example, hyperonymy is the most implemented relation). Of course the richness of different types of relations creates problems and unclear determination of these relations. In case of hyperonyms the developers of EstWN have encountered problems in choosing preferably only one suitable hyperonym for each concept. When dealing with meronymy the more specific relations – involved location, involved direction (both source and target direction) – are inconsistently determined. There are, however, no significant problems with involved instrument and involved agent relations. In PWN there is no involved location of direction relation explicitly available. Meronymy relations are often associated with the problems of connecting encyclopedic concepts to those of general language, for example how to connect the concept 'bird' to a specific bird species.

In EstWN the general language vocabulary is well covered, specific domain vocabularies are also incorporated (architecture, medicine, economy etc.) and it would be useful to connect specific vocabulary to general language vocabulary. The paper proposes that the answer to this problem could be the complementary information provided from domain labels. The last semantic relation discussed in this paper deals with fuzzynymy, since this is the third used relation in EstWN. Fuzzynymy is a free association relation, but it is clear that some groups form out of the fuzzynymy relation that can be defined as new specific relations in Estonian.

Recently EstWN has become an increasingly used resource in Estonian language technology, and as such it is important to improve the quality and consistency of relations in addition to increasing the amount of concepts in EstWN in different domains.

Keywords: lexical semantics, computational lexicology, language resources, computational linguistics, Estonian

Heili Orav
arvutiteaduse instituut
eesti ja üldkeeleteaduse instituut
Tartu Ülikool
Juhan Liivi 2
50409 Tartu
heili.orav@ut.ee

Sirli Zupping
eesti ja üldkeeleteaduse instituut
Tartu Ülikool
Jakobi 2
51014 Tartu
sirli.zupping@ut.ee

Kadri Vare
arvutiteaduse instituut
Tartu Ülikool
Juhan Liivi 2
50409 Tartu
kadri.vare@ut.ee