

EESTI WORDNET'I HETKESEISUST

HEILI ORAV, KADRI KERNER, SIRLI PARM

Eestis on erineva sisu ja ülesehitusega sõnastikke palju. Neist mõistelisti sõnastikke kaks: esimene Andrus Saareste „Eesti keele mõisteline sõnaraamat” (1958–1963) ja teine Tartu Ülikoolis koostatav uuema põlvkonna arvutitesaurus Eesti Wordnet. Viimasest selles artiklis juttu tulebki.

Artiklis anname ülevaate tesauruse senisest koostamiskäigust ja esilekerkinud probleemidest. Artikkel jaguneb kolmeks osaks: esimeses antakse ülevaade tesauruse ülesehitusest, teises on kirjeldatud tesauruse täiendamisega seotud teemasid ja kolmandas osas arutletakse selle üle, millised keeletehnoloogilised rakendused võiksid olla Eesti Wordnet'iga seotud ja kuidas lahendada esilekerkinud probleeme pikas perspektiivis.

1. Eesti Wordnet

1.1. Taustast

Tartu Ülikooli arvutilingvistika uurimisrühma semantikaalased tööd said alguse 1995. aastal, kui hakati tegema ettevalmistusi eesti üldkeele masinloetava tesauruse koostamiseks. Selle vajalikkus oli ilmne, sest Eestis oli sõnastikke palju – nii arvutisse viidud pabersõnastikke kui ka juba arvutiga koostatud sõnastikke –, kuid mitte ühtegi arvutisõnastikku, mille eesmärk ja struktuur erineksid tavasõnastikust. 1996. aastal liitus TÜ arvutilingvistika uurimisrühm Euroopa Komisjoni rahastatud projektiga EuroWordNet (www.illc.uva.nl/EuroWordNet/), mille eesmärgiks oli luua mitmekeelne wordnet-tüüpi tesaurus. Seega on meie arvutitesauruse koostamiseks koos mõne aastase pausiga kulunud üle kümne aasta.

Eesti Wordnet'i ehk TÜ eesti keele tesauruse (www.cl.ut.ee/ressursid/teksaurus/) kallal käib töö siiani. Keele leksikaal-semantiline andmebaas, kus on lisaks sõnade tähenduste eristamisele fikseeritud ka tähendustevahelelised seosed, on oluline nii lingvistiliseks uurimistööks kui ka arvutilingvistilisteks rakendusteks, nagu sisupõhine infootsing, masintõlge jms. Eri keelte wordnet-tüüpi leksikaal-semantilisi andmebaase on loodud üle viiekümne ning wordnet'id on loomuliku keele rakendustes kõige kasutatavamaid ressursse. Väärtuslikuks muudab Eesti Wordnet'i ka selle mitmekeelsus, sest EuroWordNet on mitmekeelne andmebaas, mille kaudu on Eesti Wordnet'is olevad mõisted ühendatud ingliskeelse keeltevälise indeksi abil [InterLingual Index (ILI-link)] ka teiste keelte tesauruste mõistetega.

1.2. Wordnet'i ülesehitus

Wordnet-tüüpi tesauruse eeskujuks on Princetoni Ülikooli psühholingvistide koostatud inglise keele WordNet (<http://wordnet.princeton.edu/>), mida loojad iseloomustavad leksikaalsete viidete süsteemina, mille ülesehitus põhineb

psühholingvistilistel teooriatel inimpsüühika leksikaalsest organisatsioonist ja mälust (Beckwith jt 1991; vt ka Fellbaum 1998). Seega on *wordnet* leksikaal-semantilise andmebaasi ehk arvutitesaurusena oma ülesehituselt mõistepõhine, s.t selle tesauruse põhiüksuse – sünohulga – moodustavad ühe mõiste väljendamiseks kasutatavad üks või mitu sünonüümset sõna (nt nimisõnaline mõiste: *pang*, *ämber*; tegusõnaline: *õhutama*, *agiteerima*, *kihutama*, *ässitama*, *kehutama*; või omadussõnaline: *rõõmsameelne*, *rõõmus*, *rõõmsatujuline*, *päikseline*). Üheliikmelised sünohulgad moodustavad tesauruses hetkel 57 % kõikidest sünohulkadest ning kõige rohkem on ühes sünohulgas 20 liiget.

2011. aasta jaanuariks olime jõudnud rohkem kui 42 000-mõistelise andmebaasini, milles on kasutusel 43 erinevat tüüpi semantilist seost (vt lähemalt Vider jt 2000). Esindatud on nimi-, tegu-, omadus- ja määrsõnad (vt joonist 1). Polüseemia indeks (vrd ka Langemets 2009: 76; <http://wordnet.princeton.edu/>) – arvestades ka monoseemseid sõnu – on 1,24 tähendust sõna kohta ning ainult polüseemseid sõnu arvestades 2,48 tähendust (arvestatud on ka liitsõnu). Kõige polüseemsemad on tesauruses tegusõnad ning seejärel adverbid.

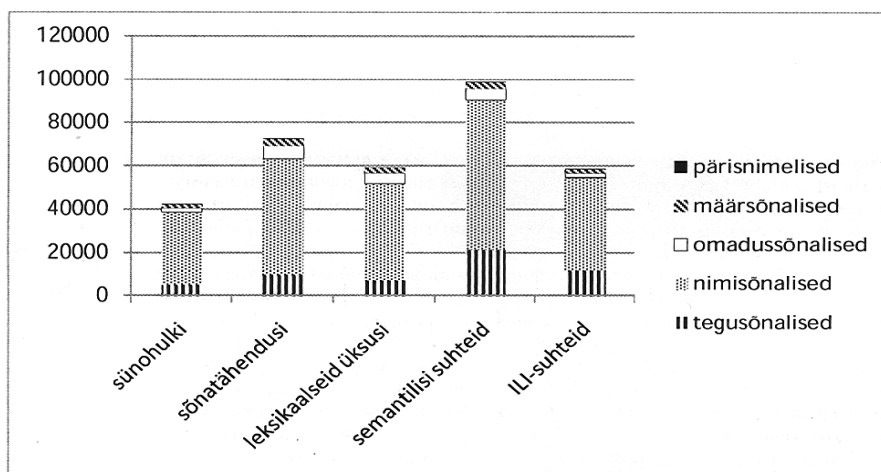
Mõisteid seovad erinevad viidad – semantilised suhted –, mis teevadki selle arvutiressursi väga oluliseks. Sellised suhted esindavad sõnade tähenduste vahelisi assotsiatsioone, mis koos moodustavad kompleksse semantilise võrgustiku: teades, kus asub sõna selles võrgus, milline on ta kontekst, suudame paremini eristada sõna tähendust. Tegusõnadel ja nimisõnadel esineb kõige rohkem hüperonüümia suhet ning omadussõnadel ja määrsõnadel (läh)antonüümia ja (läh)sünonüümia suhet.

2. Ülevaade semantilistest suhetest

Tuntuimateks semantiliste suhete teoreetilist tausta avavateks käsiraamatuteks on lingvistikas kujunenud Alan D. Cruse'i „Lexical Semantics” (1986) ja „Meaning in Language” (2000), kus antakse ülevaade erinevatest semantilistest suhetest inglise keele näidete varal. Kuid samu probleeme on käsitletud ka mitmes keele olemust üldteoreetilisest aspektist analüüsivas teoses, näiteks kontseptuaalse semantika looja Ray Jackendoffi „Foundations of Language. Brain, Meaning, Grammar, Evolution” (2002), kus läbi mitme peatüki arutletakse tähenduse olemuse üle, eriti sõnade nn puhtkeelelise tähenduse ja kontseptuaalsete teadmiste seose üle, tähendustevaheliste seoste ning erinevate sõnaliikide semantiliste iseärasuste üle.

Psühholingvistide eesmärki – ehitada sõnavõrgustik, mis jäljendaks mentaalset leksikoni – kontrolliti keelekõnelejaid testides. Idee fikseerida semantilisi suhteid nii, et seosed oleks kajastatud selliselt, nagu need toimivad inimajus tähendustest arusaamise protsessides, on iseenesest hea, kuid mõistetevaheliste viitade määratlemisel selguvad hoopiski ebaselgust tekitavad kitsaskohad. Testide tulemusena avastati, et peale sünonüümiasuhte on nimisõnade puhul mõnest mõistest arusaamiseks olulised veel alam- ja ülemmõisted, tegusõnade puhul kausaalsus ja omadussõnade puhul lähisünonüümia ning antonüümia.

Mõistetevaheliste semantiliste suhete määramisel on vaja arvestada ja ühitada kaks erinevat kriteeriumi: esiteks eesti keele sõnavara eripärad



Joonis 1. Eesti Wordnet arvudes.

suhted teoreetilises mõttes (missuguseid keele sõnavara semantilise struktuuri uusi ja olulisi aspekte on vaja ning võimalik leida) ja teiseks, tesauruse olulisust keeletehnoloogiliste rakenduste jaoks, s.t mida rohkem ja detailsemalt on suhetevõrgustik loodud, seda parema arvutiressurssi me saame. Siiani oleme järginud inglise keele vastava andmebaasi ülesehitust, mis ei toeta omakeelset lähenemist. Seega on peale tesauruse mahu suurendamise meie eesmärgiks kindlaks teha eesti keele jaoks rakendatavad semantilised suhted, mis siiani tesauruses puuduvad või on ebatäielikud. Näiteks on määrsõnade semantilist analüüsi eesti keele tesauruse jaoks tehtud alles aasta.¹ Seni tehtu põhjal saab öelda, et üks lahendust vajavaid probleeme on eesti keele tuletussüsteem, kus adverbide ja teiste sõnaliikide vahel esineb tuletusseoseid, mida praegune tesauruse formaat kajastada ei võimalda:² näiteks esineb n-ö rederivatsioon, mille puhul adverbist tuletatud adjektiivile lisatakse omakorda adverbisufiks *-lt* (adv *kohe* > adj *kohe+ne* > adv *kohe+se+lt*). Praegune tesauruse formaat võimaldab sellisel juhul ühe sünohulga ehk mõiste alla määrata erinevaid lekseeme nii, et näiteks adverbid *kohe* ja *koheselt* koonduvad samasse sünohulka ning ühendatakse tuletussuhte abil adjektiiviga *kohene*. Sellist tuletusseost võiks märkida nn tagasituletuse leksikaal-semantiline suhe. Määrsõnade erisuseks on ka see, et nende tähendusskaala on üsna lai ja neist paljud esinevad mitmes eri tähenduses, näiteks sõna *veel* märgib nii ajalist suhet (*Olen veel kodus*), hulgasuht (*Kallan kannu veel vett*) kui ka rõhutusefunktsiooni (*Kes need siis veel on!*) (adverbide tähenduste kohta vt Parm 2007: 124). Seega peaks adverbide tesaurusesse sisestamisel olema võimalik leksikaal-semantiliste suhete abil määrata ka nende tähendusega seotud baaskategooria (kas aja-, ruumi-, hulga-, seisundikategooria vms) ja/või vastavate uurimustega välja selgitada semantiline alus, mis seob selliste sõnade erinevaid tähendusi. Adverbide täpsed semantilised suhted on edasistes keeletehnoloogilistes rakendustes olulised, sest näiteks masintõlkes vajatakse

¹ EuroWordNet'is adjektiive ega adverbe ega seega ka nendega seotud semantilisi probleeme ei käsitleta.

² Sarnase probleemiga puutuvad kokku ka nt slaavi keelte tesauruste koostajad (vt nt Bosch jt 2008; Koeva jt 2008).

se adverbide tähenduste kohta kindlasti märksa rohkem infot kui infootsingus või refereerimisel.

Samuti vajab omadussõnade leksikaal-semantiline analüüs senisest märksa laiemaid ja keelepõhisemaid seoseid, näiteks on vaja siduda omavalhel teatavad konkreetset või abstraktsed objektid neid iseloomustava omadusega: KUJU → *ümmargune, kandiline*; TEMPERATUUR → *soe, külm*; ISIK-SUSEOMADUS → *õrn, maias, lahke*. Lisaks on omadussõnadel oma hierarhiasüsteem: nii on mõiste *armukade* ülemmõiste *kade* ja *külalislahke* ülemmõiste *lahke*.

Nimisõnadel vajab fikseerimist n-ö kaasnevuse suhe (näiteks kuidas on seotud *maa* ja *maamaks* või *omanik* ja *leiutasu*) ja funktsioonisuhe (näiteks *veduri* funktsiooniks on *vedada rongi*) jt. Et semantiliste suhete nn äärealasid pole eriti uuritud, siis on esitatud näited paraku juhuslikud ja ilmselt üksikud paljude hulgast. Ka teiste keelte *wordnet*'i koostajad on samas olukorras: lisaks universaalsetele semantilistele seostele tuleb leida ka oma keele spetsiifilised suhted.³

Konkreetsete sõnade suhteid on võimalik kontrollida ka Eesti Wordnet'i tähendusnumbrite põhjal sõnatähendusi ühestanud inimeste erimeelsusi uurides (vt Kerner 2004), sest reaalses tekstis võivad semantilised suhted erineda olemasolevas tesauruses esitatutest. Sõnatähenduste automaatne määramine on vahendav ülesanne, sest võimaldab teistel rakendustel paremini toimida (Resnik 2006: 299), ning kõige rohkem kasutataksegi automaatsel sõnatähenduste ühestamisel just *wordnet*-tüüpi tesauruse tähenduste järgi märgendatud korpusi.

2. Tesauruse täiendamine

Eesti Wordnet'i täiendamine toimub kahel viisil: nii kvantitatiivselt kui ka kvalitatiivselt. Esmalt kirjeldamegi sõnastiku mahu suurendamist.

Eri keelte *wordnet*'i koostajad on kasutanud arvutisõnastiku loomisel erinevaid strateegiaid, mille hulgas olulisemad on:

- täielikult käsitsi, teisi sõnastikke ja/või korpusi on kasutatud ainult konteksti kontrollimiseks (nt ingliskeelne Princetoni Ülikooli WordNet, vt <http://wordnet.princeton.edu/>);
- poolautomaatselt (nt taani ja islandi keele *wordnet*-tüüpi tesaurused);
- täielikult automaatselt (nt soomekeelne *wordnet*-tüüpi sõnastik, kus ingliskeelne Princetoni WordNet tõlgiti soome keelde, mida hiljem redigeeriti).

2.1. Sõnastiku automaatse suurendamise katse

Leksikaalselt põhineb loodav tesaurus olemasolevatel traditsioonilistel sõnaraamatutel – peamiselt „Eesti keele seletaval sõnaraamatul” (EKSS) – ja tekstikorpustel, mis annavad teavet sõna kasutamisest. Omakeelse *wordnet*'i koostamine on meil olnud peamiselt käsitsi töö, ehkki oleme proovinud tesauruse mahtu suurendada ka automaatselt. Automaatne lähenemine tähendab

³ Teiste keelte tesauruste koostamisel tekkivate semantiliste suhete probleeme kajastavad Global WordNet'i konverentsiettekannete kogumikud. Konverentse on korraldatud iga kahe aasta järel, alates 2002. aastast (vt ka <http://www.globalwordnet.org/>).

seda, et tuleb kokku panna ükskeelseid tähendusi eristavaid ja/või seletavaid ning sünonüüme esitavaid sõnastikke. Eesti keeles vastasid neile kriteeriumidele EKSS-i elektrooniline variant ja Filosoofi teaurus (http://www.filosoofi.ee/thes_et/). Viimane on omakorda automaatselt kokku pandud Asta Öimu „Sünonüümisõnastikust” (Öim 1991) ja „Antonüümisõnastikust” (Öim 1995).

Probleemidest, mis tekkisid tesaurust automaatsel viisil täiendades, kaitsi 2009. aasta kevadel bakalaureusetöö (Villem 2009). Nimisõnade puhul täheldasime automaatse täiendamise korral peamiselt järgmisi probleeme.

1. Saime palju sünohulki, mis olid ebakorrektsed. Esines nii tähenduste üleeristamist, mille puhul kahe tähenduse vaheline erinevus on väga ähmane ja sageli vaid kontekstist sõltuv, kui ka alaeristamist, mille puhul seletuses esineb rinnastav *või*, mis viitab tegelikult kahele erinevale tähendusele (Vider jt 2000: 137). Näiteks moodustusid Filosoofi tesaurusest automaatselt järgmised sünohulgad: *jaanalind*, *emu*, kus tegemist on erinevate linnuliikidega; *šampus*, *vahuvein* veini alammõistetena, ehkki on teada, et need ei ole üks ja seesama; *karupüksid*, *retuusid*, *liibukad*, kus *karupüksid* saab öelda ainult villast kootud pükste kohta. Toodud näidete põhjal saab öelda, et kvaliteetses *wordnet*-tüüpi tesauruses peaksid säilima (liigi)spetsiifilised mõisted.

2. Saime palju sünohulki, mille puhul jäi silma ainsuse ja mitmuse küsimus. Näiteks oli sünonüümiridadesse kokku pandud *rinnad*, *tissid*, *rinnapartii* või *kaerahelbed*, *herkulo*.

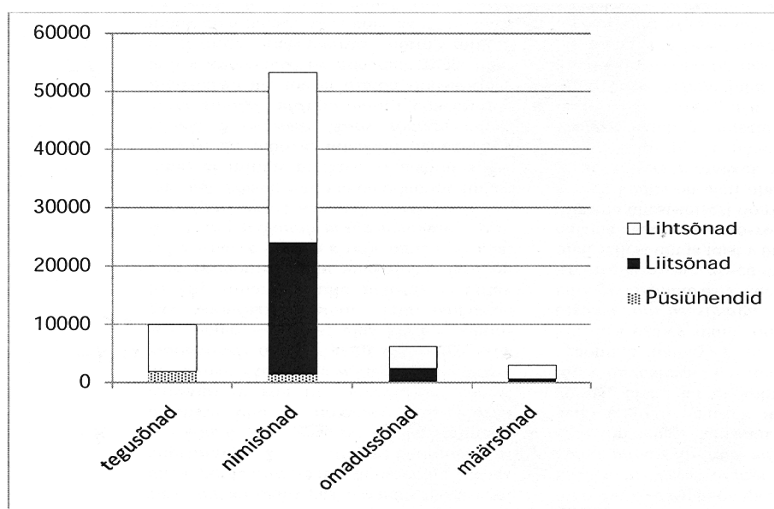
3. Saime sünohulki, mida mõjutab sooline eripära. Näiteks *klassikaaslane*, *klassiõde*, *klassivend*, kus *klassikaaslane* peaks olema *klassiõe* ja *klassivenna* ülemmõiste.

Tegime järelduse, et eestikeelse tesauruse puhul peaksime jätkama peamiselt käsitsi tööga, sest automaatselt ülekantud sünohulkade kontrollimisele kulus liiga palju aega. Kontrollimise käigus kustutati palju automaatselt ülekantud sünohulki või sünohulkade liikmeid ning palju sünohulki liideti omavahel. Mõistete käsitsi lisamisel tavaliselt selliseid probleeme ei teki.

2.2. Liitsõnad ja püsiühendid

Eesti keeles on väga palju – kui mitte öelda lõputu hulk – liitsõnu ja püsiühendeid. Näiteks olevat liitsõnade osakaal „Eesti keele seletavas sõnaraamatus” (22 esimese vihiku põhjal) umbes 61 % (Langemets 2004). Tesauruse koostajatel tuleb otsustada, mil määral ja kuidas liitsõnu tesaurusesse lülitada (vt nt Lõo 2010). Probleem pole mõistagi uus, iga leksikoni koostaja on liitsõnadega hädas olnud, kuid nagu öeldud, täieneb meie sõnastik suuresti sõnatähenduste ühestamise tulemustest. Palju uusi mõisteid ja puuduvaid tähendusi lisandub sõnatähenduste ühestamise projektist⁴ ja sealt lisandub palju just reaalses tekstis kasutatavaid liitsõnu. Näiteks ilukirjandustekstide ühestamise tulemusena saame poeetilisi, kunstilisi ja mitteüldkeelseid liitsõnu – nii peaksid tesaurusesse sattuma ka näiteks sellised sõnad nagu *õrnusevaru*, *õõnetuvi* jms. ESTMORF-i morfoloogiline analüsaator (<http://www.eki.ee/keele tehnoloogia/projektid/estmorf/>) tuvastas tesauruses üle 14 000 liitsõna (vt joonist 2), mis moodustab umbes 30 % kõikidest sõnadest. Liitsõnade lisamisel tuleb jälgida, kas tegemist on piisavalt sagedase ja üldkeeles tuntud liitsõnaga.

⁴ Riikliku programmi „Eesti keele keeletehnoloogiline tugi” projekt „Eesti keele semantika ressursid ja vahendid” (vt ka <http://www.keeletehnoloogia.ee>).



Joonis 2. Lihtsõnade, liitsõnade ja püsiühendite arv sõnaliigiti Eesti Wordnet'is.

Samalaadne on ka püsiühendite küsimus. Kui idioome käsitleda lihtsalt pikkade sõnadena, mille vorm ja tähendus on nii-öelda kivistunud, saab neid lisada *wordnet*-tüüpi arvutisõnastiku sõnohulkadesse valmiskujul. Seega võib Eesti Wordnet'i andmebaasi lisada püsiühendid *kägu ajama* ja *hambasse puhuma* sõnohulka *valetama* ja püsiühend *keelt peksma* sõnohulka *klatsima*. Mainitud idioomid on vaid üheti mõistetavad ja küllalt laialt levinud, niisiis oleks nende kaasamine mõttekas. Ometi ei ole võimalik kõiki idioome käsitleda niisama lihtsalt, sest nende sõnastikku lisamisel võib tekkida mitmesuguseid süntaktilisi, semantilisi ja muidki takistusi. Näiteks eitavas vormis väljendid *mitte sõrmeotsagagi puutuma*, *ei küsi leiba*, *ei löö risti ette* kaotavad jaatavas kõnes oma õige tähenduse.

Väljendid *härjal sarvist haarama* ja *muresid pudelisse uputama* pole siiski sõnonüümsed mõistetega *tegutsema* ja *jooma* ning kannavad endas spetsiifilisemat tähendust. *Tahab olla igas pulmas pruut ja igal matusel surnu* inimese iseloomujoone tähistajana on tesauruse sisendiks ilmselgelt liiga pikk ja sellele pole leida ka ühtegi (lähi)sõnonüümi.

Verbikeske püsiühendi puhul kontrollitakse esmalt, kas see on olemas püsiühendite andmebaasis (vt lähemalt Muischnek 2006) ning seejärel otsustatakse (sagedust ja sobivust silmas pidades) tesaurusesse lisamine. Praeguse seisuga on Eesti Wordnet'is üle 2800 püsiühendi (vt joonist 2) ning püsiühendid on esindatud kõikides sõnaliikides.

2.3. Taksonoomiate kontroll

Eesti Wordnet'i loomisel on peamisi eeldusi see, et sõnavara põhituumik võiks olla ja peaks olema organiseeritud taksonoomiatena. Püüame hoolitseda, et vähemalt nimi- ja tegusõnalistel mõistetel oleks fikseeritud ülem- ja alamõiste. Teised sõnaliigid (nagu omadus- ja määrsõnad), mille puhul on keeruleine täiuslikku taksonoomilist lähenemist rakendada, peavad olema seotud kas lähisõnonüümia- või (lähi)antonüümiasuhtega.

Nagu öeldud, käib tesauruse mahu suurendamine käsikäes olemasolevate mõistete ja nende semantiliste seoste täpsustamisega. Üks ülevaatamist vajavaid osi ongi tesauruse taksonoomiad, mille moodustavad hierarhilised suhted hüponüümia ja hüperonüümia.

Hüponüümia tähistatakse tähenduste hierarhilisi alistussuhteid. Alammõiste sõna on oma ülemmõiste suhtes hüponüüm, ülemmõiste sõna oma alammõiste suhtes hüperonüüm. Hierarhias ülalpool on väga suuremahulised üldised mõisted, allapoole eritunnuste hulk kasvab ja mõiste maht väheneb.

Testides seda suhet (vt Cruse 1986: 137), saame hüperonüumiat eristada nii:

X on teatav Y
või kui see on X, siis peab ta olema ka Y

Ehk siis, kui ta on *alkohoolik, kodutu, jalakäija, eestlane, iludus* jne, siis peab ta olema ka *inimene*.

Eesti Wordnet'i põhjal on Riin Kirt oma bakalaureusetöös uurinud inimesega seonduvat hierarhiat (Kirt 2009). Eesmärk oli Eesti Wordnet'is oleva mõiste *inimene* hüponüümide kategoriseerimine ning nende omavaheliste semantiliste suhete kontrollimine. Töö käigus selgus, et sünohulga *inimene, inimolend, indiviid, isik, persoon, hingeline, hing* alla on lisatud põhjendamatult palju alammõisteid – üle 700.

Mõiste *inimene* otseste hüponüümide arvu vähendamiseks tuleks inimesega seonduv sõnavara jaotada nii (vt Kirt 2009):

- ametite järgi, nt *assistent, poesell, sepp, käskjalg, jahimees*;
- tegevuse järgi, nt *organisaator, osavõtja, sportlane, reisija, põgenik*;
- isikuomaduste järgi, nt *mängur, väljapressija, pätt, elunautija, õnnekütt, kääbus*;
- (emotsionaalsete) seisundite järgi, nt *möku, õnnetu, haige, prominent, noor*;
- asukoha järgi, nt *elanik, välismaalane, eestlane, külaline, pärismaalane*;
- (lähi)suhete järgi, nt *alluv, tööandja, vaenlane, ketser, lähedane, armsam*.

Kinnitust sai, et on vaja üle vaadata ka teiste valdkondade hierarhiaid, sest see aitab olemasolevat süsteemi korrastada ning loob hea aluspõhja uutele lisatavatele mõistetele.

2.4. Valdkonnamärgendite lisamine

Hetkel tegeldakse ka tesauruse liigendamise ja süstematiseerimise seisukohalt sõnavara valdkondliku esituse põhimõtete ja praeguse olukorra selgitamisega. Selleks on esialgu valitud keelest sõltumatud *wordnet*-tüüpi leksi-kaal-semantilistele andmebaasidele mõeldud märgendid (vt Magnini, Cavaglià 2000). Märgendite lisamisel tekiks omavahel seotud mõistete pered ehk semantilise välja kaudu seotud mõistekogumid, millel on vastav märke juures. Näiteks *kool* seotakse *õppeasutuse* väljaga, *traktor põllumajandusega* jne. Ka teiste keelte *wordnet*'id kasutavad neidsamu valdkonnamärgendeid, näiteks on kogu sõnavara liigendatud valdkonniti projektis MultiWordNet (<http://multiwordnet.itc.it/english/home.php>), kus valdkonnamärgendid kantakse automaatselt üle Princetoni Ülikooli WordNet'ist itaalia ja hispaania keele *wordnet*-tüüpi tesaurusesse.

Neid valdkonnamärgendeid on võimalik Eesti Wordnet'i andmebaasi lisada poolautomaatselt, kusjuures märgendid täiendavad juba olemasolevat

informatsiooni: iga sünohulk saab vähemalt ühe lisamärgendi umbes kahe-sajast hierarhiliselt struktureeritud märgendist. Valdkonnamärgendid on nimisõnadele, tegusõnadele, omadussõnadele ja määrsõnadele ning neid võib olla rohkem kui üks. Sama valdkonnamärgendi võivad saada eri sõnaliigid. Näiteks eeltoodud näitega sarnaselt saaks inimese valdkonnamärgendi nii *liiderlik*, *seelikukütt* kui ka *flirtima*. Valdkonnamärgendite abil võib rühmitada sõnatähendusi, mille tulemusena väheneks ka polüseemiatase. Üks *wordnet*'i puudusi ongi liiga üleeristatud sõnatähendused, mis on peamine probleem ka automaatse ühestamise juures (McCarthy 2006: 17). Valdkonnainformatsiooni kasutatakse ka sõnatähenduste ühestamisel (Buitelaar jt 2006: 288).

3. Tesauruse tulevikuülesanded

Oluline on märkida, et meie tesauruse algne idee – katta eesti keele üldkeele sõnavara – on praeguseks paratamatult jõudnud erikeelse lähenemiseni, s.t tesauruse laiendamise töö toimub teatavate valdkondade sõnavara kaupa ja ületab kaugelt üldkeele piirid, minnes kohati väga spetsiifiliseks. Nii on põhjalikult läbi töötatud näiteks transpordivahendite (Raat 2007), ilmastikunähtuste (Konsap 2008), isiksuseomaduste (Orav 2006) sõnavara. Käsil on arhitektuuri, filosoofia, kirjandusteaduse sõnavara. Selline valdkonnaspetsiifiline lähenemine muudab mõistete ja nende semantiliste suhete lisamise täpsemaks ja kergemaks, sest korraga tegeleb ühe valdkonna mõistetega üks inimene. Samuti saab eri keelte *wordnet*'i-tüüpi tesaurusi leksikaal-semantilisest seisukohast kõrvutada ja võrrelda mitmesuguste semantiliste väljade struktuuri. Projektis EuroWordNet paluti kõigil osalejatel peensusteni välja arendada näiteks muusikariistade, mõõteriistade ja emotsioonide hierarhiad, mida võrreldi keeltevälise indeksi abil leksikaliseerituse, hüponüümiatasandite ja paljude muude näitajate poolest (Vider jt 2000).

Peamiselt oleme tegelnud käsitsi tesauruse suurendamisega, kuid alustasime juba ka tuletiste poolautomaatse lisamisega, nimelt *mine*-liiteliste nimisõnade *ma*-tegevusnimest tuletamisest, sest tegusõnad on tesauruses peaaegu täielikult esindatud, aga *mine*-liitelisi nimisõnu ei leidu peaaegu üldse. *mine*-liide on eesti keeles väga produktiivne ning muudab harva tähenduse sisu. Samuti on võimalik üle kanda olemasolevad semantilised suhted. Regulaarne on ka tegijanimede tuletus. Teiste sõnaliikide puhul on automaatselt tuletamine keerulisem, kuid edaspidi on plaanis tegelda omadussõnade ja määrsõnade tuletistega: produktiivse *lt*-liite abil saab omadussõnadest tuletada suure hulga viisiadverbe. Et selline sõnamoodustusviis on lihtne ja läbi-paistev, siis on võimalik eesti keele tesauruse mahtu kõvasti suurendada neid määrsõnu automaatselt moodustades.

Viimasel ajal on ka keeletehnoloogias aktuaalseks muutunud ontoloogia-te loomine ja kasutamine (vt nt <http://www.ontologyportal.org>). Arvutiteaduses on ontoloogia kontseptualisatsiooni eksplitsiitne (ilmutatud) ja formaalne spetsifikatsioon (Gruber 1993). *Wordnet*-tüüpi leksikone peetakse taksonoomiateks, ka lingvistilisteks ontoloogiateks, kuid siiski mitte päris ontoloogiaks. *Wordnet*-tüüpi leksikonid on piiratud vaid mingi keele sõnade ja sõnaühenditega, seal pole metalingvistilist informatsiooni (Schalley, Zaefferer 2007: 7). Küll aga on neist võimalik saada ontoloogiatega koostamiseks väärtuslikku informatsiooni. Ning ka vastupidi: *wordnet*-tüüpi tesaurusi on võimalik

ontoloogiatega põhjal täiendada ja parandada. Näiteks on Princetoni WordNet'i põhjal (leksikaalselt) täiendatud edukalt inimese anatoomia ontoloogiat (Reiter, Buitelaar 2008: 378).

Ka riikliku programmi „Eesti keele keeletehnoloogiline tugi 2009” projekti „Nutika süvaveebi- ja veebiressursse kombineeriva infootsisüsteemi prototüübi” raames (vt <http://ats.cs.ut.ee/semantika/wiki/index.php/Projektist>) luuakse eesti üldontoloogia ja valdkonnaontoloogiaid ning üldontoloogia loomiseks kasutatakse ka juba olemasolevat ressursi Eesti Wordnet.

Artikli valmimist on toetanud sihtfinantseeritavad teadusteemad „Keel ja tähendus: semantika ja grammatika kognitiivses perspektiivis” ja „Loomulike keelte arvutitöötamise formalismide ja efektiivsete algoritmide väljatöötamine ning eesti keelele rakendamine” ning Euroopa Regionaalarengu Fond, Eesti arvutiteaduse tippkeskus ja riikliku programmi „Eesti keele keeletehnoloogiline tugi” projekt „TÜ eesti keele tesauruse (Eesti Wordnet'i) täiendamine”.

Kirjandus

- Beckwith, Richard, Fellbaum, Christiane, Gross, Derek, Miller, Georg 1991. WordNet. A Lexical Database Organized on Psycholinguistic Principles. – U. Zernik (toim), Lexical Acquisition. Exploiting On-line Resources to Build a Lexicon. Hillsdale, NJ: Erlbaum, lk 211–231.
- Bosch, Sonja, Fellbaum, Christiane, Pala, Karel 2008. Enhancing WordNets with Morphological Relations. A Case Study from Czech, English and Zulu. – A. Tanacs, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (toim), Proceedings of the Fourth Global WordNet Conference. Szeged, Hungary, January 22–25, 2008. Szeged: University of Szeged, lk 74–90.
- Buitelaar, Paul, Magnini, Bernardo, Strapparava, Carlo, Vossen, Piek 2006. Domain-Specific WSD 2006. – E. Agirre, P. Edmonds (toim), Word Sense Disambiguation. Algorithms and Applications. (Text, Speech and Language Technology, Vol. 33.) Netherlands: Springer, lk 275–298.
- Cruise, D. Alain 1986. Lexical Semantics. Cambridge: Cambridge University Press.
- Cruise, D. Alain 2000. Meaning in Language. An Introduction to Semantics and Pragmatics. Oxford: Oxford University Press.
- EKSS = Eesti keele seletav sõnaraamat. 1–6. Tallinn: Eesti Keele Sihtasutus, 2009 (vt <http://www.eki.ee/dict/ekss/>).
- Fellbaum, Christiane 1998. An WordNet, Electronic Lexical Database. Toim Ch. Fellbaum. Cambridge–Massachusetts: The MIT Press.
- Gruber, Thomas R. 1993. A Translation Approach to Portable Ontology Specifications. – Knowledge Acquisition, nr 5, lk 199–220.
- Jackendoff, Ray 2002. Foundations of Language. Brain, Meaning, Grammar, Evolution. New York: Oxford University Press.
- Kerner, Kadri 2004. Sõnatähtsused tekstides ja tesauruses ühestajate erimeelsuste põhjal. (Bakalaureusetöö, Tartu Ülikool, üldkeeleteaduse õppetool).
- Kirt, Riin 2009. Inimesega seotud hierarhiapuu Eesti Wordnetis. (Bakalaureusetöö. Tartu Ülikooli eesti ja üldkeeleteaduse instituut).
- Koeva, Svetla, Krstev, Cvetana, Vitas, Duško 2008. Morpho-semantic Relations in WordNet – a Case Study for Two Slavic Languages. – A. Tanacs, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (toim), Proceedings of the

- Fourth Global WordNet Conference. Szeged, Hungary, January 22–25, 2008. Szeged: University of Szeged, lk 239–253.
- K o n s a p, Gaili 2008. Ilmastikunähtuste sõnavara leksikaal-semantiline analüüs eesti keeles. (Bakalaureusetöö, Tartu Ülikool, eesti ja üldkeeleteaduse instituut).
- L a n g e m e t s, Margit 2004. Polüseemia ja leksikograafia. – Emakeele Seltsi aastaraamat 49, 2003. Tallinn: Eesti Teaduste Akadeemia, Emakeele Selts, lk 97–124 (http://www.emakeeleselts.ee/esa/ESA_49_pdf/Langemets.pdf).
- L a n g e m e t s, Margit 2009. Nimisõna süstemaatiline polüseemia eesti keeles ja selle esitus eesti keelevaras. Doktoritöö. Tallinn: Tallinna Ülikool.
- L õ o, Kaidi 2010. Püsiühendid ja liitsõnad *wordnet*-tüüpi tesauruses. (Bakalaureusetöö, Tartu Ülikool, eesti ja üldkeeleteaduse instituut).
- M c C a r t h y, Diana 2006. Relating WordNet Senses for Word Sense Disambiguation. – Proceedings of the ACL Workshop on Making Sense of Sense: Bringing Psycholinguistics and Computational Linguistics Together. Trento, lk 17–24.
- M a g n i n i, Bernardo, C a v a g l i à, Gabriela 2000. Integrating Subject Field Codes into WordNet. – Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May–2 June. Athens, lk 1413–1418.
- M u i s c h n e k, Kadri 2006. Verbi ja noomeni püsiühendid eesti keeles. (Dissertationes philologiae Estonicae Universitatis Tartuensis 17.) Tartu: Tartu Ülikooli Kirjastus.
- O r a v, Heili 2006. Isiksuseomaduste sõnavara semantika eesti keeles. (Dissertationes linguisticae Universitatis Tartuensis 6.) Tartu: Tartu Ülikooli Kirjastus.
- P a r m, Sirli 2007. Partiklite *veel*, *juba*, *alles*, *jälle* tähendused eesti kirjakeeles. (Magistritöö, Tartu Ülikool, üldkeeleteaduse õppetool).
- R a a t, Alvin 2007. Transpordi ja transpordivahenditega seotud mõistete leksikaal-semantiline analüüs. (Bakalaureusetöö, Tartu Ülikool, eesti ja üldkeeleteaduse instituut).
- R e i t e r, Nils, B u i t e l a a r, Paul 2008. Lexical Enrichment of a Human Anatomy Ontology Using WordNet. – A. Tanacs, D. Csendes, V. Vincze, Ch. Fellbaum, P. Vossen (toim), Proceedings of the Fourth Global WordNet Conference. Szeged, Hungary, January 22–25, 2008. Szeged: University of Szeged, lk 375–386.
- R e s n i k, Philip 2006. WSD in NLP Applications. – E. Agirre, P. Edmonds (toim), Word Sense Disambiguation. Algorithms and Applications. (Text, Speech and Language Technology, Vol. 33.) Netherlands: Springer, lk 299–326.
- S a a r e s t e, Andrus. 1958–1963. Eesti keele mõisteline sõnaraamat I–IV. Dictionnaire analogique de la Estonienne I–IV. Stockholm: Vaba Eesti.
- S c h a l l e y, Andrea C., Z a e f f e r r, Dietmar 2007. Ontolinguistics. – A. C. Shalley, D. Zaeffeerr (toim), An Outline. Ontolinguistics. How Ontological Status Shapes the Linguistic Coding of Concepts. (Trends in Linguistics. Studies and Monographs 176.) Berlin–New York: Mouton de Gruyter, lk 3–23.
- V i d e r, Kadri, K a h u s k, Neeme, P a l d r e, Leho, Õ i m, Haldur, O r a v, Heili 2000. Eesti keele teaurus. – T. Hennoste (toim), Arvutuslingvistikalt inimesele. (Tartu Ülikooli üldkeeleteaduse õppetooli toimetised 1.) Tartu: Tartu Ülikooli Kirjastus, lk 127–152.
- V i l l e m, Olga-Anniki 2009. ILI-kirjete lisamine Eesti wordnetti ja selle käigus ilmnunud automaatselt genereeritud sünohulkade probleemkohad. (Bakalaureusetöö, Tartu Ülikool, eesti ja üldkeeleteaduse instituut).
- Õ i m, Asta 1991. Sünonüümisõnastik. Tallinn.
- Õ i m, Asta 1995. Antonüümisõnastik. Tallinn: Keele ja Kirjanduse Instituut.

Estonian Wordnet Today

Keywords: Estonian Wordnet, wordnet, computational lexicography, human language technology

This article describes the creation of the Estonian Wordnet and discusses the main problems that have been dealt with during recent years. At present the Estonian Wordnet consists of more than 46,000 lexical units, which form more than 30,000 concepts (synsets). The Estonian Wordnet, at the present stage, includes nouns, verbs, adjectives and adverbs.

The concepts in the Estonian Wordnet are connected by 43 different types of semantic relations. While determining these relations there are two main criteria to consider. First, there is a need for Estonian language specific relations and, second, there is a need (considering language technological applications) for a rich network of semantic relations. This paper presents a list of Estonian language specific relations which should be present in the Estonian Wordnet, while the main focus is on adjectives and adverbs.

So far, our approach to enlargement has mainly been manual and domain-specific, i.e. we have gradually added semantic fields such as architecture, transportation, personality traits. There was also an attempt to enlarge Estonian Wordnet semi-automatically by transferring around 3000 new noun synsets from the Estonian Synonym and Antonym Dictionary. It turned out, however, that there were too many synsets which had to be corrected, revised, joined together or deleted, and the work of revising took up too much time.

Also, this paper describes the problems of adding multiword units and compound words into the Estonian Wordnet, since the Estonian language has an infinite capacity for compounding. It was found that it is important to add the most frequent ones. Another ongoing work consists in the inclusion of domain labels from WordNet Domains.

Besides the enlargement of the Estonian Wordnet we have started revising the existing data. One of the problems is the revision of hierarchies and so far one study has been carried out – the checking of the taxonomy of 'human being' in the EstWN. Our future plans include automatic addition of synsets generated by derivation, since Estonian is rich in derivatives. We will start from suffixes that are highly regular, and move on to cases requiring morphological analysis and synthesis.

Heili Orav (b. 1969), PhD, University of Tartu, researcher, Institute of Estonian and General Linguistics, heili.orav@ut.ee

Kadri Kerner (b. 1981), doctoral student, University of Tartu, Institute of Estonian and General Linguistics, kadri.kerner@ut.ee

Sirli Parm (b. 1981), doctoral student, University of Tartu, Institute of Estonian and General Linguistics, sirli.parm@ut.ee