TARTU ÜLIKOOL
FILOSOOFIATEADUSKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT
ARVUTILINGVISTIKA ERIALA

Kaidi Lõo

Püsiühendid ja liitsõnad *wordnet*-tüüpi tesauruses

Bakalaureusetöö

Juhendaja Heili Orav, Ph.D.

TARTU 2010

Sisukord

Sissejuhatus

Käesoleva töö eesmärgiks on uurida püsiühendite ja liitsõnade kui keeles väga levinud nähtuste lisamist *wordnet*-tüüpi tesaurusesse. Siiani ei ole olemas kindlat teoreetilist ega praktilist alust, kuidas püsiühendeid ja liitsõnu *wordnet*-tüüpi tesaurustesse lisada ning ka töö Eesti Wordnet'is on olnud selles valdkonnas küllaltki juhuslik.

Töö teoreetilises osas tuuakse ära põhimõistestik – mis on püsiühendid, mis liitsõnad, kuidas on nende omavaheline suhe, mida mõeldakse arvutisõnastike ning mida täpsemalt WordNet'i ja mida Eesti Wordnet'i all.

Töö praktilise osa moodustab esmalt ülevaade, millised probleemid on tekkinud püsiühenditega seoses ingliskeelses WordNet'is, tuues samas ka eestikeelseid analooge. Teiseks analüüsitakse sagedasemaid eestikeelseid verbikeskseid püsiühendeid ja kolmandaks liitsõnu Eesti Wordnet'is. Kokkuvõttes püütakse jõuda mõnede lahenduste pakkumiseni, mida peaksid Eesti Wordnet'i tegijad püsiühendite lisamisel tesaurusesse silmas pidama.

1. Püsiühendid

Püsiühend (ingl *fixed expression*, *multi word expression*, *multi-word unit*) on kahe või enama sõna(vormi) ühend, mida mingi tähenduse väljendamiseks on tavaks koos kasutada (Muischnek 2006: 12).

Siia hulka kuuluvad mitmesugused ilmingud keeles – näiteks idiomaatilised üksused, liitsõnad, terminoloogia, pärisnimed, verbi partikli konstruktsioonid, tugiverbiühendid, metafoorid jne (Sag jt 2001).

Püsiühend kordub tavaliselt enam-vähem ühel ja samal kujul suur arv kordi, võimaldab suhelda aja- ja energiasäästlikumalt ning lisab keelele kujundlikkust (Krikmann 2004: 103).

Eesti keeleteaduses nimetatakse püsiühendite ehk fraseemide uurimisega tegelevat valdkonda fraseoloogiaks. Neid püsiühendeid, millele on omane osade tähenduslik kokkukuulumine ja harilikult ka metafoorsus, nimetatakse fraseologismideks. (EKK 2000: 520)

Eesti keele käsiraamatus (EKK 2000: 520) on püsiühendid jagatud kaheks kollokatsioonid ja idioomid. Kollokatsioon on sõna tähendusest sõltuv kalduvus esineda koos kindlate teiste sõnadega. Mõned sõnad kollotseeruvad omavahel, teised mitte. Kõige selgemini väljenduvad kollokatsioonid selles, missuguse aluse, sihitise, määruse saab enda juurde valida öeldiseks olev verb. Nt *määgib* puhul saab aluseks olla *lammas*, mitte *kala* või *pliiats.* Kollokatsioonid on ka näiteks *silmi kissitama*, *küsimust esitama*, *õlgu kehitama*. (EKK 2000: 520)

Idioom on tähenduslikult kokkusulanud liikmetega püsiühend, mille kogutähendus ei tulene ühendit moodustavate sõnade tähenduste summas. Näiteks *südant puistama*, *villast viskama*. (EKK 2000: 520)

Kadri Muischnek (2006) on oma doktoritöös idioomid ja kollokatsioonid veel omakorda jaotanud rühmadesse. Idioomid on jagatud läbipaistmatuteks ja läbipaistvateks vastavalt sellele, kas keelekasutaja mõistab ühendeid ilma neid eraldi omandamata. Läbipaistmatud idioomid on näiteks *jalga laskma*, *lugu pidama* ja läbipaistvad *põhja laskma*, *sadulas püsima*. Kollokatsioonid on jällegi jaotatud pool-

idioomideks ja tugiverbiühenditeks. Pool-idioome võib vaadelda kui idioomide ja kollokatsioonide ühisosa, tugiverbiühendid võivad oma kombineerumisomaduste poolest olla kas kollokatiivsed, näiteks *kõnet pidama* või vabad ühendid, näiteks *eksamit sooritama*. (Muischnek 2006: 26)

Üks kuulsamaid ja mõjukaimaid idioomide käsitlusi on toodud ka Nunbergi ja tema kaasautorite artiklis „*Idioms*" ajakirjas *Language* (Nunberg jt 1994) Artiklis on idioom määratletud kui hägus kategooria ja seda on kirjeldatud prototüüpsete tunnuste abil. Ainus kõigile idioomidele kehtiv tunnus on konventsionaalsus ehk kokkuleppelisus. See tähendab, et idioomi tähendust või kasutust ei saa järeldada või vähemalt täielikult järeldada kokkulepetest, mis määravad tema üksikute osade tähenduse. (Nunberg jt 1994: 492)

Teised tunnused, mida mainitakse, on samuti tüüpilised idioomidele, kuid ei kehti alati. Idioomidele on näiteks tihti iseloomulik süntaktiline jäikus – nad esinevad vaid piiratud arvus süntaktilises raamides ja tarindites. Idioomid on tavaliselt ka kujundlikud ja sisaldavad metafoore, hüperboole ja teisi kõnekujundlikke väljendeid. Nad on rahvalikud, see tähendab, et neid kasutatakse, et kirjeldada ja selgitada sarnasuse või seotuse kaudu koduste ja konkreetsete asjadega mingit korduvat sotsiaalset situatsiooni. Enamasti on idioomid ka kõnekeelsed ja neid kasutatakse mitteametlikus ja suulises kõnes. Idioomidele on omane emotsionaalsus, nad väljendavad hinnangulist või afektiivset suhtumist asjadesse, mida nad tähistavad. Keel ei kasuta tavaliselt idioome, et kirjeldada neutraalseid olukordi. (Nunberg jt 1994: 492-493)

Eelnevalt oli näha, et kuigi keeleteaduses on idioomide uurimisega tegeletud päris palju, puudub sellele keelendile ikkagi üks täpne määratlus.

2. Liitsõnad

Liitmine ehk kompositsioon on selline sõnamoodustusviis, kus uue sisulise ja vormilise terviku ehk liitsõna annab kahe, mõnikord ka rohkema tüve ühendamine. Näiteks *aed + linn = aedlinn.* Liitsõnad võivad sisemiselt olla mitmesuguse moodustusstruktuuriga. Nii sõnaliigi kui ka muuttüübi määrab liitsõna viimane moodustusosa – põhiosa. Erinevalt põhiosast jääb eesmine moodustusosa, täiendosa, üldiselt kõigis liitsõna muudetes muutumatuks. (EKG I 1995: 411)

Liitsõnamoodustus tähendab võimalust nimetada asju, nähtusi, tegevusi ja omadusi mõne iseloomuliku tunnuse või seose põhjal. Seos, millest liitsõna moodustamisel lähtutakse võib olla vahetu, kuid ka kaudne, see tähendab ülekandeline. (EKG I 1995: 411)

Liitmisel on oluline osa eriti nimisõnamoodustuses ja see on eesti keeles väga produktiivne sõnamoodustusviis. Sõltuvalt sellest, milline on moodustusosade tähenduste vahekord ja millist rolli täidavad nad liitsõna terviktähenduse seisukohalt, eristuvad determinatiivsed ja kopulatiivsed liitnimisõnad. Erinevalt determinatiivsetest liitnimetustest (näiteks *paberkott* – põhiosa *kott* märgib liitsõna põhilist sisu, täiendosa ülesandeks on seda täpsustada) ei väljenda kopulatiivsete liitnimisõnade (näiteks *ööpäev* märgib kokku ööd ja päeva) järelkomponent üksi nimetusega väljendatu üldmõistet, vaid nimetuse sisu koosneb mõlema, teineteise suhtes võrdse kaaluga moodustusosa tähenduste summast. (EKG I 1995: 476)

Fraseoloogilised liitsõnad on tõenäoliselt suures osas determinatiivsed. Selle põhjuseks võib pidada asjaolu, et taoliste liitsõnade moodustusosad ei ole sugugi võrdse kaaluga. Liitsõna osiste vahelisele osutavast tähendusest sõltub see, millises ulatuses on liitsõna tähendus motiveeritud, see tähendab moodustusosadega määratud. (Baran 2004: 159)

Eesti keeles kirjutatakse liitsõna osad alati kokku, ükskõik mis vormis või ümbruses nad ka ei esineks. Mõnel juhul võib liitsõnades ka sidekriipsu panna, kuid see ei ole kohustuslik, seda näiteks parema loetavuse ja mõtteselguse huvides (*korvpalli-*

meistrivõistlused) või siis liitsõna piiril, kus satuvad kõrvuti kolm või enam ühesugust tähte (*maa-ala*). (EKK 2000: 270)

Eesti keeles on suur hulk selliseid täismetafoorseid liitsõnu, mida võib pidada fraseoloogilisteks liitsõnadeks, näiteks *hädapätakas*, *laiskvorst* jne. Kuigi fraseologismiks peetakse tavaliselt kahest või enamast sõnast koosnevat ühendit, ei ole koostisosade hulk otsustava tähendusega. Pealegi, nagu teada, on taoliste liitsõnade teke pelgalt ortograafia probleem ja kokkuleppe asi. (Õim 1993: 23)

Inglise keeles võib liitsõnu kirjutada kolme moodi: 1) kokku - näiteks *keyboard* (eesti keeles *klaviatuur*), *newspaper* (eesti keeles *ajaleht*) 2) sidekriipsuga - näiteks *mother-in-law* (eesti keeles *ämm*), *world-wide* (eesti keeles *ülemaailmne*) 3) lahku - näiteks *compound word* (eesti keeles *liitsõna*), *match box* (eesti keeles *tikutoos*) (Carter, McCarthy 2006: 321)

Järelikult võivad nii inglise keele kui teatud mööndustega ka eesti keele liitsõnad kuuluda püsiühendite alla.

Edasi annan vastavalt töö eesmärgile lühikese ülevaate arvutisõnastikest, mille keelematerjalist suure osa moodustavad just püsiühendid.

3. Arvutileksikon

Arvutuslingvistilised rakendussüsteemid, näiteks infootsingu-, keeleõppe-, tõlkesüsteemid, ei toimi piisava leksikonita, kuid see leksikon ei saa olla ainult sõnade vorme fikseeriv (analüüsiv, sünteesiv) süsteem, vaid peab sisaldama ka piisavalt semantilist ja pragmaatilist informatsiooni. Seetõttu on leksikonide koostamine tänapäeva arvutuslingvistiliste rakendussüsteemide jaoks väga oluline. (Orav, Vider 2006: 85)

Arvutisõnastik ei ole vaid arvutisse viidud sõnaraamatu tekst. Sõnastikuartikli erinevad funktsionaalsed osad (märksõna ise, grammatiline info, seletus, näited) peavad olema formaalselt identifitseeritavad ja olema varustatud spetsiifiliste märgenditega. Just tänu sellisele liigendusele on sõnastikus esindatud materjal ka „arvuti pool loetav“ (*machine readable dictionary*) ning mitte ainult raamatu asendaja. Arvuti abil võidakse otsida või analüüsida eraldi sõnastikuartikli erinevaid osi, näiteks seletusi, näiteid ja grammatilist infot. (Orav, Vider 2006: 86)

1980. aastatel leiti, et on vaja andmebaasi vormi, mis oleks kasulik automaatselt taksonoomiate, seletuste jms tegemiseks. Arvutisõnastikke hakati kasutama erinevate semantiliste hierarhiate ehitamiseks. Võtmesõnaks sai leksikaal-semantiline andmebaas. Tuntuim selline on semantiline arvutileksikon WordNet. (Orav, Vider 2006: 87)

3.1. WordNet

1985. aastal otsustas rühm Princetoni ülikooli psühholingviste eesotsas Georg Milleriga luua inglise sõnavara leksikaalse andmebaasi, mis erinevalt varasematest ei oleks üles ehitatud mitte tähestikuliselt, vaid mõisteliselt. See põhimõte määras ka käsitletavate sõnade hulga. Andmebaasi märksõnadeks said olla vaid täistähenduslikud sõnad: nimi-, arv-, tegu-, omadus- ja määrsõnad. Et senised psühholingvistilised uurimused, millele tugineti, näitlikustasid oma hüpoteese peamiselt sõnaassotsiatiivtestidega, mis hõlmasid sadakonda ingliskeelset sõna, oli üks WordNet'i loomise motiive katsetada neid samu hüpoteese suuremat hulka üldsõnavara kasutades.

Töö tulemusena selgus, et süntaktiliste kategooriate (sõnaliikide) põhilised erinevused on selgelt nähtavad ja kasutusel ka nende semantikas. Nii on nimisõnad leksikaalses mälus korrastunud tipnevate hierarhiatena, paljud omadussõnad kui *n*-mõõtmelised „hüperruumid“ ja tegusõnad on omavahel mitmekesistes järelduvussuhetes. (Orav, Vider 1998: 58)

Sünohulk ehk sünonüümirida (ingl *synset*, *synonym set*) on WordNet'i elementaarosake, mille moodustavad ühte mõistet (tähendust) väljendavad sünonüümsed sõnad (ja ka sõnaühendid). Termin *sünohulk* on loodud sellepärast, et erinevalt sünonüümisõnastiku sünonüümireast võib WordNet'is sünohulk olla ka üheliikmeline. Kui sünonüümisõnastiku eesmärgiks on kõigi võimalike keeles leiduvate sünonüümide esitamine, siis WordNet'is on mõisted esitatud ka siis, kui selle väljendamiseks keeles leidub ainult üks leksikaalne üksus.[1]

Princeton WordNet'i viimane versioon 3.0 sisaldab umbes 157 000 sõna, mis on organiseeritud üle 115 000 sünohulka. Sõnadest ligi 120 000 moodustavad nimi- ja arvsõnad, 21 000 omadussõnad, ligi 11 500 tegusõnad ja 4500 määrsõnad.[2]

WordNet'is ühendavad sünohulki erinevad semantilised seosed. Peamiselt hüpo- ja hüperonüümia, antonüümia, osa-terviku suhted, põhjuslikkus- ja rollisuhted, kuid ühtekokku on WordNet'is esindatud ligi 60 erinevat suhtetüüpi. (Orav, Vider 2006: 87)

Inglisekeelne WordNet on olnud eeskujuks mitmekümne teise *wordnet*-tüüpi tesauruse ja rakenduste loomisele. *The Global WordNet Association* kooridineerib ja juhib uute *wordnet*- tüüpi tesauruste väljatöötamist ning korraldab iga kahe aasta tagant rahvusvahelisi konverentse.[3]

[1] http://www.cl.ut.ee/ressursid/teksaurus/ (22.05.2010)
[2] http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html (22.05.2010)
[3] http://wordnet.princeton.edu/ (22.05.2010)

3.2. Eesti Wordnet ehk Tartu Ülikooli eesti keele tesaurus

1998. aastast alates on Tartu Ülikooli arvutilingvistika uurimisrühmas koostatud eesti üldkeele tesaurust ehk kogu sõnavara haarata püüdvat suurt sõnaraamatut, mis koos viidetega ingliskeelsele WordNet'ile moodustab Eesti Wordnet'i ja on üks kaheksast EuroWordNet-2 projekti tulemusena saadud ja ELRA kaudu levitatavast *wordnet*-tüüpi tesaurusest. Eesti Wordnet'i peamised tegijad on olnud Neeme Kahusk, Heili Orav, Kadri Vider jt ning järgitud on Princetoni WordNet'i ja EuroWordNet'i põhimõtteid.[4]

Hetkel on Eesti Wordnet'is (aprill 2010) sünohulki ligikaudu 29 000, sellest umbes 20 000 nimisõna, 5 000 tegusõna, 2000 omadussõna, 1500 määrsõna ja 600 pärisnime. Üle poolte sünohulkadest on üheliikmelised, veidi üle veerandi on kaheliikmelisi ja 3- kuni 9-liimelisi sünohulki on 19% kõigist sünohulkadest. Kõigist tesauruses esitatud leksikaalsetest üksustest (sõnadest ja sõnaühenditest) 80% on esindatud ühe tähendusega. Mitmetähenduslikest (kaks ja enamat tähendust) üksustest moodustavad omakorda 68 % kahe tähendusega esindatud sõnad, kusjuures verbid on polüseemsemad kui ülejäänud sõnaliigid.[5]

Semantilisi seoseid on ühel sünohulgal üle kahe, domineerivad hüpo- ja hüperonüümiasuhted, kuid kasutusel on 43 erinevat semantilise suhte tüüpi. [6]

Eesti keele tesauruse koostamine jätkub igapäevaselt ja selle käigus ei pääse selle tegijad otsustamisest, kuidas võiksid püsiühendid ja liitsõnad olla esindatud eestikeelses arvutisõnastikus.

[4] http://www.cl.ut.ee/ressursid/teksaurus/ (22.05.2010)
[5] http://www.cl.ut.ee/ressursid/teksaurus/ (22.05.2010)
[6] http://www.cl.ut.ee/ressursid/teksaurus/ (22.05.2010)

4. Püsiühendid ja *wordnet*-tüüpi tesaurus

4.1. Püsiühendite lisamise vajalikkus

Nagu eespool kirjeldatud on püsiühend mitmekesine ja hägune mõiste, millega on põhjalikumalt tegelema hakatud alles viimaste kümnendite jooksul, seetõttu puudub siiani ka ühtne arusaam, mis kuulub täpselt püsiühendite alla ja kuidas neid liigitada. Nende laia kasutuse tõttu sõnavaras oleks aga kindlasti kasulik lisada püsiühendeid rohkem ka klassikalistesse pabersõnaraamatutesse, kuid eriti just nende lisamine arvutipõhistesse *wordnet*-tüüpi tesaurustesse peaks parandama püsiühendite paindlikumat analüüsi. (Fellbaum 2006: 350)

Birte Lönneker ja Carina Eilts toovad oma artiklis „*A Current Resource and Future Perspectives for Enriching WordNets with Metaphor Information*“ (2004) välja, miks oleks süstemaatiline metafoorse info lisamine kasulik. Nimelt aitaks see mitmeid erinevaid keeletehnoloogia rakendusi, mis kasutavad *wordnet*-tüüpi tesaurusi oma töös, näiteks info-otsing, sõnatähenduste ühestamine ja keeleõpe.

Info-otsing võidaks palju, kui metafoorsed tähendused oleksid selgelt eristatud, kuna alati ei kehti tähenduste puhul paralleelne polüseemia. Kui leksikaalsed ressursid nagu WordNet oleksid varustatud metafoorsete tähenduste ja seletustega, areneks kiiremini ka sõnatähenduste ühestamine ja see võimaldaks automaatselt luua semantiliselt märgendatud korpuse masintõlke jaoks. On tõestatud, et mõisteliselt struktureeritud sõnade nimestikud aitavad inimestel sõnavara paremini meeles pidada, niisiis paraneks metafoorsete tähenduste lisamisel oluliselt ka keeleõppe tase. (Eilts, Lönneker 2004:2)

Alljärgnevalt kirjeldan mõningaid takistusi, mis on tekkinud inglisekeelse WordNet'i loojatel seoses idioomide ehk ühe püsiühendite alaliigi lisamisega *wordnet*-tüüpi tesaurusesse. Et probleemkohti paremini selgitada, lisan ka eestikeelseid näiteid.

4.2. Probleemid idioomide lisamisel *wordnet*-tüüpi tesaurusesse

Kui idioome käsitleda lihtsalt pikkade sõnadena, mille vorm ja tähendus on nii-öelda kivistunud, ei ole nende lisamine arvutisõnastikku WordNet kuigi raske, sest neid saab lisada sünohulkadesse juba valmis kujul. Näiteks on läbipaistmatud inglisekeelsed idioomid *kick the bucket* (ülekantud tähenduses *surema*) ja *buy the farm* (samuti ülekantud tähenduses *surema*) mõlemad sünohulga *die* (eesti keeles *surema*) liikmed. (Fellbaum 1998: 53)

Eesti Wordnet'i andmebaasi võib analoogselt lisada näiteks *kägu ajama* ja *hambasse puhuma* sünohulka *valetama* või siis *keelt peksma* sünohulka *klatšima*. Mainitud idioomid on vaid üheti mõistetavad ja keeles küllalt laialt levinud, niisiis oleks nende lisamine vajalik.

Ometi ei ole võimalik kõiki idioome käsitleda sama lihtsalt, sest nende lisamisel sõnastikku võib tekkida kõikvõimalikke süntaktilisi, semantilisi kui ka muid takistusi. Alljärgnevalt tähtsamatel neist peatungi.

4.2.1. Süntaktilised probleemid

Siia hulka kuuluvad probleemid, mis tekivad, kuna mõnede idioomistringide pindvormid ei ühildu ühegi WordNet'i süntaktilise kategooriaga. Näiteks saavad paljud idioomid esineda ainult eituses ja kaotavad muidu oma tähenduse. Väljendeid *not give a hoot* (eesti keeles *mitte hoolima*), *not hold a candle* (eesti keeles *kõrvale jätma*) on WordNet'is keerukas käsitleda, sest nende verbifraaside peasõna on eituses. (Fellbaum 1998:54)

Ka mitmed eesti keele verbifraasid, nagu *mitte sõrmeotsaga puudutama*, *ei küsi leiba*, *ei löö risti ette* kaotavad nagu eelpool mainitud inglisekeelsed näitedki jaatavas kõneliigis oma õige tähenduse.

Eitus ole aga probleemiks, kui taolised idioomid on sünonüümsed juba mõne WordNet'is oleva stringiga. Näiteks *not in a pig's eye*, *when hell freezes over* ja *when the cows come home* (kõik kolm on idiomaatilised väljendid, mille tähendus on *mitte kunagi*) on kõik sünonüümsed stringiga *never* (eesti keeles *mitte kunagi*), mis on juba

WordNet'i koosseisus määrsõnade all. Seetõttu võib eelpool mainitud idioomid lisada lihtsalt samasse sünohulka. (Fellbaum 1998:54)

Eesti Wordnet'is saaks analoogselt lisada *mitte lillegi liigutama*, mis on sünonüümne stringiga *laisklema* ja *mitte piuksugi tegema*, mis on sünonüümne stringiga *vaikima*.

Teine süntaktiline probleem on see, et paljude verbifraasiliste idioomide osad ei ole idioomides järjest ja nendes on muutuvaid osi. Inglise keeles on mitmeid selliseid idioome, kus genitiivis nimi- või asesõna on subjektiga seotud. Näiteks *flip one's wig* (eesti keeles *marru sattuma), blow one's stack* (eesti keeles *vihaseks saama*), *cook one's goose* (eesti keeles *kellegi võimalusi rikkuma*). Selliseid idioome ei saa vaadelda kui ühte stringi, sest *one* eelmistest näidetest võib asendada mistahes nimi- või asesõnaga. Üks lahendus oleks lisada sellised idioomid nii, et puuduv sõna asendada teatud metamärgiga. Selline kirje oleks aga raskesti töödeldav. Kui aga süntaktilisel märgendamisel saab metamärgi asemele sõnu lisada vaid kindlast nimi- ja asesõnade hulgast, oleks tulemus parem. (Fellbaum 1998:54)

Ka eesti keeles on taolisi muutuvate liikmetega idioome. Näiteks mitmed alalütleva käände ja *olema*-verbiga fraasid: *kellelgi on sussid püsti*, *kellelgi on vaim peal, kellelgi on jänes püksis* jt

Niisiis ei saa eelmainitud idioome lisada, kui nende süntaktiline vorm ei vasta ühelegi WordNet'i kategooriale, kuid probleemi ei teki, kui sellistel idioomidel on WordNet'i korraldusega kokkusobivaid sünonüümseid stringe. (Fellbaum 1998:55)

4.2.2. Semantilised probleemid

Lisaks süntaktilistele probleemidele tekib idioomide lisamisel ka mõistelis-semantilisi takistusi. Mõned idioomid väljendavad mõisteid, mis ei sobitu andmebaasi ei sünohulkade liikmetena ega ka iseseisvate mõistetena, kuna WordNet'is lihtsalt pole teisi selliseid leksikaliseerunud mõisteid, millega nad võiksid seoses olla. Nii selgub fraseoloogiasõnaraamatuid lugedes kiiresti, et idioomid esindavad tihti keerukaid ja mitmetasandilisi mõisteid, mida ei saagi ümber sõnastada tavaliste leksikaalsete või süntaktiliste kategooriate abil. (Fellbaum 1998:55)

Näiteks ei anna inglisekeelse idioomi *drown one's sorrows* tähendust täielikult edasi tegusõna *drink* (eesti keeles *jooma*) või idioomi *fish bait* tähendust verbiühend *take action* (eesti keeles *samme astuma*). Taolised idioomid kannavad endas palju sellist eriomast semantilist informatsiooni, mis läheks tõenäoliselt kaduma, kui nad lisada WordNet'i üldisemate mõistete alla. (Fellbaum 1998:55)

Eesti keeles on vastavaid näiteid väga palju, näiteks *härjal sarvist haarama* ja *muresid pudelisse uputama* ei ole päris sünonüümsed sõnaga *tegutsema* ja *jooma* ning kannavad endas üldisemat tähendust.

Mõnel juhul paraneks idioomi tähenduse edasiandmine, kui see tähendus osadeks jaotada ja neile osadele omakorda tähendused anda. Näiteks *spill the beans* (eesti keeles *saladust avaldama*) tähenduse võib osadeks jaotada järgmiselt. *Spill* vastab tähendusele *reveal (*eesti keeles *avaldama* ja *beans* tähendusele *secret* (eesti keeles *saladus*) . (Fellbaum ja Osherson 2010:3)

Ka eesti keeles on taolisi näiteid. Idioomis *kuivale jooksma*, mis tähendab *raskustesse sattuma*, on *kuiv* sünonüümne tähendusega *raskus* ja *jooksma* tähendusega *sattuma.* Niisiis võiks tesaurusesse idioomi lisamisel ja suhete seadmisel teiste sünohulkadega seda arvestada.

Teine lahendus idioomide semantilistele probleemide lahendamiseks oleks eraldi uue sünohulga loomine, mis sisaldaks idioomi komponenti. Enamikul sellistel juhtudel oleks sünohulgas küll vaid üks liige, kuid mõnel juhul viitavad erinevad sõnad erinevates idioomides samale mõistele. Näiteks on *cat* (eesti keeles *kass*) ja *beans* (eesti keeles *oad*) idioomides *let the cat out of the bag* ja *spill the beans* (eesti keeles *saladust avaldama)* mõlemad *saladuse* tähenduses. (Fellbaum, Osherson 2010: 3)

Eestikeelsetes idioomides *pada ajama* ja *umbluud rääkima* on *pada* ja *umbluu* sama metafoorse sünohulga liikmed tähenduses *loba*.

4.2.3. Eessõnad

Hetkel ei kuulu eessõnad WordNet'is kodeeritud sõnaliikide hulka ning seetõttu ei ole WordNet suuteline veel idioomile spetsiifilisi eessõnu eraldi töötlema. Põhjuseks on nende ebaselge staatus, mis on kusagil täistähendusliku ja abisõna vahel. Tihti on aga eessõnadel idioomide tähenduses tähtis roll. (Osherson, Fellbaum 2010: 4)

Näiteks väljend *sweep something under the carpet* (eesti keeles *midagi vaiba alla pühkima*) tähendab ülekantud tähenduses *midagi varjama*, kuid väljendil *sweep something on the carpet* on vaid otsene tähendus ehk siis *midagi vaiba peale pühkima*. Seetõttu peaksidki WordNet'i tegijad töötama selle kallal, et tulevikus oleks võimalik idioomidele spetsiifilisi eessõnu paremini käsitleda. (Osherson, Fellbaum 2010: 4)

Kuigi eesti keeles väljendavad eessõnade tähendust tavaliselt käände lõpud, on mõnikord ka eesti keeles õige eessõna valik oluline. Väljendi *midagi vaka alla panema* metafoorne tähendus on *oma tõelisi võimeid varjama*. Väljendil *midagi vaka sisse panema* on aga vaid otsene tähendus ehk siis *midagi teatud kujuga nõusse asetama.* Niisiis on võib valesti töötlemisel tekkida oluline tähenduserinevus.

4.2.4 Keerulised lausekonstruktsioonid

Mõningaid mitmesõnalisi üksusi pole võimalik ega mõttekas WordNet'i andmebaasi lisada, sest nad varieeruvad selleks nii süntaktiliselt kui ka semantiliselt liiga laialt. Näiteks *the cat's got your tongue* (eesti keeles *vaikima sellisel juhul kui rääkimine oleks tegelikult vajalik*) või *eat one's cake and have it too* (eesti keeles *tahtma kahte asja, mida tegelikult korraga omade ei saa*). Neil väljenditel on mitmeid erinevaid tähendusvarjundeid ning ka süntaktiliselt esineb varieerumist. Näiteks võib viimasena mainitud idioom olla kujul *have one's cake and eat it too* või *have one's cake and eat it* ja tähendada lisaks ka *saama midagi, mida ei olda ära teeninud*. (Fellbaum 1998: 54)

Ka eesti keeles on palju näiteid sellistest keerulistest üksustest, mida ei ole mõtet eesti keele tesaurusesse lisada. Näiteks *ennast kas või ribadeks naerma , tahab olla igas pulmas pruut ja igal matusel surnu*, *kergemat vastupanu teed minema* jne Neid ei ole saa tesaurusesse lisada, sest neis võib esineda nii suuri süntaktilisi kui ka semantilisi kõrvalekaldeid.

5. Verbikesksed püsiühendid Eesti Wordnet'is

Siiani on püsiühendite lisamine Eesti Wordnet'is olnud küllaltki juhuslik ja toimunud põhiliselt siis, kui neil on olnud sünonüüme. Et analüüsida, kuidas on püsiühendid hetkel Eesti Wordnet'is esindatud, võtsin tekstikorpuses vähemalt 10 korda esinevate verbikesksete püsiühendite loendi (vt Tabel 1.) ning kontrollisin kõigepealt millised vastavatest püsiühenditest on juba Eesti Wordnet'is (TEKSauruses) olemas, kas nad on õigete semantiliste suhetega ning milliseid sealt veel puuduvad püsiühenditest võiks tesaurusesse veel lisada ja kuidas. Abimaterjalina kasutasin analüüsimisel „Eesti kirjakeele sõnaraamatut"[7] ja „Sünonüümisõnastikku"[8].

[7] http://www.eki.ee/dict/ekss/ (25.05.2010)
[8] http://dict.ibs.ee/syn/ (25.05.2010)

Tabel 1. Tekstikorpuses vähemalt 10 korda esinevad verbikesksed püsiühendid (Muischnek 2006: 83)

Püsiühend	**Esinemine TEKSauruses**
aru saama	+
end tundma	-
meelde tuletama	+
tööd tegema	+
võimalust andma	+
toime tulema	+
kätte saama	+
tähele panema	+
selgeks saama	+
ettepanekut tegema	+
meelde tulema	+
endast kujutama	-
huvi tundma	+
tähelepanu pöörama	-
kasutusele võtma	-
küsimustele vastama	-
silmas pidama	+
õlgu kehitama	-
aega võtma	-
vaeva nägema	+
mõju avaldama	+
kirja panema	-
endale lubama	-
aset leidma	+
arvesse võtma	+
pähe tulema	+
muljet jätma	-
selgeks tegema	+
paika panema	+
osa võtma	+
kindlaks tegema	+
andeks andma	+
silma paistma	+
otsa vaatama	-
käsku andma	+
algust saama	+

Kontrolli käigus selgus, et enamik eelneva tabeli püsiühenditest on küll eesti keele tesaurusesse lisatud, kuid esineb mõningaid puudusi ja ebatäpsusi nende sünonüümide ja teistes semantilistes suhetes.

Esiteks ei ole sageli esindatud kõik sõna tähendused. Püsiühendil *kätte saama* pole näiteks tähendust *üles otsima*. Ka väljendverbile *tähele panema* võiks lisada teise tähenduse. Praegu esineb see tesauruses vaid tähenduses *märkama*, kuid sõna tähendab ka *jälgima*, *vaatlema*. Verbiühend *paika panema* on esindatud vaid tähenduses *midagi kindlaks määrama*, kuid puuduvad teised tähendused. Näiteks *kellelegi tema õiget kohta kätte näitama* või siis otsene tähendus *midagi õigele kohale seadma*. Väljendverb *osa võtma* on hetkel tesauruses tähenduses *osalema*, kuid sõna võib kasutada ka *kaasa tundma* sünonüümina. Näiteks *ta võttis osa sõbra õnnest, rõõmust, muredest, leinast, kurbusest.* Püsiühendile *algust saama* lisaksin samuti teise tähenduse. Nimelt peale sellele, et algust saama tähendab *lähtuma* või *pärinema*, on sel ka otsesem tähendus, nimelt *ajaliselt või kohaliselt peale hakkama*, *algama*.

Teiseks on paljudel püsiühenditel mitmed olulised sünonüümid puudu. Näiteks sünohulgas *aru saama* võiksid olla ka *aduma* ja *sotti saama*, väljendverbil *võimalust andma* puuduvad sünonüümid *lubama* ja *laskma* ning väljendverbil *selgeks saama* sünonüümid *ilmnema*, *klaaruma*, *koitma*, *lahenema*, *selginema*.

Mõnel juhul võiks mõistetele lisada ka antonüüme. Näiteks püsiühendi *tööd tegema* antonüümiks on *laisklema*, *meelde tulema* antonüümiks *ununema* ning *toime tulema* antonüümiks *läbi kukkuma*.

Üldiselt oleks võimalik peaaegu kõigile vaadeldavatele püsiühenditele juurde lisada hüponüüme ehk alammõisteid, kuid mõnel juhul oleks hüponüümide lisamine tesaurusesse sõna tähendusest arusaamise huvides eriti vajalik. Väljendverbil *huvi tundma* on hetkel vaid üks hüponüüm *silma heitma*, kuid samas kasutatakse seda väljendit tegelikult keeles väga laialt. Nii lisaksin *huvi tundma* hüponüümidena veel *harrastastama*, *fännama*, *kiinduma*, *kurameerima*, mis kõik toovad välja selle mõiste tähenduse eri tahud. Ka väljendit *tööd tegema* kasutatakse keeles väga erinevates kontekstides, niisiis lisaksin talle alammõistetena veel *teenima*, *õppima*, *vaeva nägema*, *rassima*, *pingutama*.

Kuna sõnatähenduse mõistmine on suhteline ja see sõltub oluliselt keelekasutajast, on raske välja tuua konkreetseid vigu sünohulkades või semantilistes suhetes. Nii ei teeks ma ühtegi parandust püsiühendite *arvesse võtma*, *aset leidma*, *andeks paluma* juures. Küll aga kustutaksin Eesti Wordnet'is *osa võtma* sünohulgast *partitsipeerima*, kuna ei leidnud seda ei „Eesti õigekeelsussõnaraamatust" ega ka „Eesti keele seletavast sõnaraamatust".

Nendest verbikesksetest püsiühenditest, mida hetkel veel Eesti Wordnet'is pole, jätaksingi lisamata *end tunda*, *endast kujutama*, *endale lubama*, sest need ühendid pronoomeniga ei anna minu arust sõnadele uut tähendust. Samuti ei näe ma mõtet lisada ühendeid *küsimustele vastama*, *otsa vaatama* ja *õlgu kehitama* tesaurusesse, kuna tegemist on kollokatsioonidega ehk sagedasti koosesinevate sõnapaaridega, mitte aga eraldi väljakujunenud mõistetega. Verbiühendi *tähelepanu pöörama* lisaksin aga *hoolima* sünonüümina, *kasutusele võtma* sõna *rakendama* sünonüümina, *kirja panema* on sünonüümne *märkima* ning *muljet jätma* ligikaudu sünonüümne sõnaga *vapustama*.

Niisiis võib kokkuvõttes öelda, et kuigi paljud verbikesksetest püsiühenditest on Eesti Wordnet'is olemas, tuleks seal rohkem tähelepanu pöörata erinevatele tähendustele ja lisada juurde sünonüüme, antonüüme ja hüponüüme.

6. Liitsõnad *wordnet*-tüüpi tesauruses

Liitsõnad moodustavad eesti keeles suure osa inimese leksikonist, mida näitab ka näiteks asjaolu, et liitsõnade osakaal on „Eesti kirjakeele seletavas sõnaraamatus“ umbes 61% (Langemets 2004 :102).

Järelikult peaks liitsõnu rohkem ja süstemaatilisemalt lisama ka *wordnet*-tüüpi tesaurustesse. Kui 2002. aasta *The Global WordNet Association'*i konverentsil tõdeti, et *wordnet*-tüüpi tesaurused ei sisalda peaaegu üldse liitsõnu, siis viimastel aastatel on hakatud sellega tegelema ning tesauruste tegijad on aru saanud, et kõikide mõistete sisu ja tähendust ei saa väljendada üksikute sõnadega ja seetõttu oleks vaja kindlasti lisada ka liitsõnu. (Sharada, Girish 2004: 314)

Eesti keeles – aga ka näiteks saksa keeles – on loovalt võimalik moodustada lõputul hulgal uusi liitsõnu nii, et ka teistel inimestel pole raskusi nendest arusaamisega. Näiteks on „Eesti keele seletavas sõnaraamatus“ sõna *töö* all ligi 200 liitsõna. Küsimus, mis siinkohal seoses *wordnet*-tüüpi tesaurusega tõstatub, on – milliste kriteeriumite järgi otsustada, kas mingit liitsõna oleks vaja tesaurusesse lisada või mitte.

Et analüüsida liitsõnu Eesti Wordnet'is, võtsin „Eesti keele sagedussõnastikust“ (Kaalep, Muischnek 2002) kümme sagedasemat nimisõna. Uurisin vaid neid sõnu, mis esinesid korpuses ainult nimisõnana.

Kontrollin nendest moodustavate liitsõnade esinemist ja lisamise vajalikkust Eesti Wordnet´is. Vaadeldavad liitsõnad sain „Eesti keele seletava sõnaraamatust“ (EKSS) vastavate märksõnade alt.

10 sagedasemat nimisõna eesti keeles on: aasta, mees, aeg, inimene, sõna, naine, käsi, päev, kroon, laps.

1) aasta

Praegu on EKSS-i loendis *aasta* liitsõnadest Eesti Wordnet'is juba olemas *kalendriaasta*, *liigaasta, eelarveaasta*, *kirikuaasta*, *õppeaasta*, *eluaasta*. Kindlasti võiks lisada veel *sünniaasta*, *poolaasta*, *valgusaasta*, *uusaasta*, kuna tegemist on keeles väga sageli kasutatavate mõistetega. Lisada pole aga mõtet näiteks selliseid keerulisi liitsõnu nagu *maopaoaasta*, *lisapäeva-aasta*, *õpipoisiaasta* jt, sest nad sisaldavad liiga spetsiifilisi mõisteid, mida ei kasutata keeles kuigi palju.

2) mees

Sõna *mees* on sõnamoodustuses väga produktiivne ning ka EKSS-s on mehel palju erinevaid alammõisteid, mida kõiki ei ole kindlasti mõtet tesaurusesse lisada. Näiteks on keeles vananenud mõisted nagu *tunnismees*, mis tähendab *tunnistajat* või *sulasmees*, mis tähendab *sulast*. Paljud *mehe* alammõisted ongi erinevad ametinimetused ja tähtsamad neist, nagu *ärimees*, *põllumees*, *meremees* jt, on juba tesauruse koosseisus olemas. Juurde lisaksin aga näiteks mõisted *turvamees* ja *tüürimees*, samuti sellised liitsõnad, kus *mees* ei määra sõnatähendust. Näiteks *vikatimees* tähendab piltlikult *surma*, *ninamees* piltlikult *juhti* ja *viiemees* tähendab kõnekeeles *headele hinnetele õppijat*. Levinumad otsesed *mees* alammõisteid on tesauruses üldiselt esindatud. Näiteks *noormees*, *vanamees*, *poissmees* jt

3) aeg

Mõiste *aeg* on Eesti Wordnet'is olemas kuues tähenduses ning seetõttu on seal ka küllaltki palju *aja* alammõisteid. EKSS-i nimekirjast võiks tesaurusesse veel lisada mõned sporditerminid, nagu *lisaaeg*, *poolaeg*, *võiduaeg* ning mõisted *jooksuaeg*, *ooteaeg, vaba aeg*, *kehtivusaeg*. Mõtet pole aga lisada selliseid sõnu nagu *hädaaeg*, *marjaaeg*, *paganusaeg*, kuna neid keeles enam peaaegu ei kasutatagi.

4) inimene

EKSS-i liitsõnadest on hetkel Eesti Wordnet'is järgmised nimisõnad: *omainimene*, *kunstiinimene*, *linnainimene*, *maainimene*, *mõistuseinimene*, *seltskonnainimene*, *külainimene*, *tööinimene*, *vaimuinimene*, *vanainimene*. Nimekirjast võiks juurde lisada antonüümina *mõistuseinimesele* ka *tundeinimese*, lisaks sõnad *ürginimene*, *lihtinimene*, *üliinimene* ja *lumeinimene*. Lisada pole minu arust vaja sõnu *kääbusinimene*, *naisinimene*, *meesinimene*, sest sõna *inimene* ei anna nende sõnade tähendusele midagi juurde.

5) sõna

Mõistel *sõna* on Eesti Wordnet'is üks tähendus – keeleline üksus, mida keelt emakeelena valdajad suudavad eristada. Minu arvates võiks juures olla ka veel teine tähendus. See on *sõna* laiem tähendus vestluse või teksti osana. *Sõna* esimese tähenduse enamik alammõisteid on Eesti Wordnet'is esindatud, nagu näiteks *oskus-*, *abi-*, *asesõna* jne, küll aga võiks lisada hüponüüme juurde *sõna* teisele tähendusele. Praegu on Eesti Wordnet'ist puudu *võtmesõna*, *lõppsõna*, kuid on olemas on näiteks mõisted *lööksõna*, *võlusõna*, *sõimusõna*.

6) naine

Võrreldes EKSS-i liitsõnu Eesti Wordnet'iga selgus, et tähtsamad sõnaga *naine* liitsõnad on seal esindatud ja pole mõtet lisada vähem kasutatavaid mõisteid nagu *lellenaine*, *liignaine*, *sohinaine* jt

7) käsi

Hetkel pole Eesti Wordnet'is ühtegi sellist liitsõna, mille põhisõnaks oleks *käsi.* Leian, et pole mõtet lisada ka mõistet *labakäsi*, kuna *laba* on seal juba olemas. Küll aga võiks Eesti Wordnet'is olla kaks piltlikku liitsõna, mis sisaldavad sõna *kätt*. Nimelt väljendid *kunstikäsi* ja *meistrikäsi*, mis on mõlemad ligikaudselt sünonüümid omadussõnale *meisterlik.*

8) päev

Päev on mõiste, mille sisu on Eesti Wordnet'is kaheksas eri tähenduses küllaltki täpselt esindatud. Ka enamik tähtsamaid liitsõnu EKSS-st on esindatud Eesti Wordnet'is. Lisada võiks veel mõned levinumad mõisted, nagu *nimepäev*, *südapäev*, *haigusepäev*, *puhkusepäev* ja *hällipäev*. Välja jätaksin enamik kalendritähtpäevade nimetustest, sest nimekiri EKSS-s on väga pikk ja see ei oleks ka Eesti Wordnet'i kontekstis vajalik.

9) kroon

Lisaksin Eesti Wordnet'i koosseisu liitsõna *hambakroon*, kuna see tooks välja veel ühe mõiste *kroon* tähenduse, ülejäänud *kroon* liitsõnade lisamine ei ole minu arust tarvilik.

10) laps

Kõik põhilised laps alammõisted on Eesti Wordnet'is juba olemas. Lisaksin tesaurusesse veel kaks piltlikku väljendit, mis ei ole otseselt sõna *laps* alammõisted. Nimelt *pailaps* tähenduses *soosik*, *lemmik* ning *imelaps* tähenduses *haruldaselt andekas*.

Analüüsi tulemusel võib öelda, et üldiselt on Eesti Wordnet'i koosseisus juba küllaltki palju liitsõnu, esineb aga ebatäpsusi nende suhetes ülemmõistetega ja puudu oli mitmeid keeles levinud fraseologismilisi liitsõnu ning ka muid liitsõnu.

Praegu tuleb palju uusi liitsõnu tesaurusesse sõnatähenduste ühestamise tulemusena. Kuna ühestatakse aga peamiselt ilukirjandustekste, on ka liitsõnad nendes kunstilised ja mitteüldkeelsed – nii peaksid tesaurusesse sattuma ka näiteks *õrnusevaru, õõnetuvi* jms.

See, et sõnastikke täiendatakse automaatselt korpustest, on tavaline trend keeleressursside puhul. Nii suurendavad näiteks ka tšehhid oma *wordnet*-tüüpi tesaurust liitsõnadega, mis on automaatselt korpustest välja võetud[9]. Erinevus aga tšehhi ja eestikeelsete liitsõnade juures seisneb selles, nagu eespoolgi öeldud, et liitsõnade moodustamine eesti keeles pole kindlate piiridega. Kui kirjanik on kasutanud oma

[9] Info pärineb doktorant Kadri Kerneri isiklikust vestlusest prof. Karel Palaga LREC 2010 konverentsil Maltal.

tekstis väga kujundlikku omaloomingut liitsõnade näol, siis nende lisamine tesaurusesse pole mõttekas.

Niisiis võiks tesaurusesse lisada ainult keeles levinumad determinatiivsed liitnimisõnad, näiteks *sünniaasta*, *nimepäev*, *lõppsõna* jne, sest nende koguhulk keeles on väga suur ja ka liitumisviisid küllaltki kirevad. Sellest olulisem oleks aga lisada kopulatiivseid liitnimisõnu, näiteks *poolaeg*, *valgusaasta*, sest neis on mõlemad sõnaosad võrdväärsed ja sisu moodustab nende poolt tähistatud mõistete summana. Hea oleks samuti kui tesauruses oleks rohkem keeles levinuid fraseoloogilisi liitnimisõnu, näiteks *vikatimees*, *meistrikäsi*, *pailaps*.

Kokkuvõte

Antud töö eesmärgiks oli uurida püsiühendite ja liitsõnade sisestamisega seotud probleeme *wordnet*-tüüpi tesaurustesse.

Püsiühend on mitme sõna ühend, mis koos moodustab ühtse tähendusliku terviku. Püsiühendi hägune definitsioon teeb raskeks tema eristamise teistest keelenditest ning ka liigitamise püsiühendite endi seas.

Liitsõna on selline sõnamoodustusviis, kus kahe või enama tüve liitmine annab uue sisulise ja vormilise terviku. Kuna liitsõna tähendust ei saa alati tuletada tema osade summast, võib ka liitsõnu tegelikult liigitada idiomaatiliste väljendite ehk siis laiemalt püsiühendite hulka. Traditsiooniliselt on püsiühendid aga siiski ainult mitmesõnalised väljendid.

Püsiühendite ja liitsõnade lisamise *wordnet*-tüüpi tesaurusesse teeb raskeks mõlema keelendi mitmekesisus. Püsiühendite lisamisel tesaurusesse on märgata järgmisi takistusi:

- esiteks süntaktilised probleemid, mis tulenevad püsiühendite süntaktlisest varieerumisest,
- teiseks semantilised probleemid, mis tulenevad nende ainulaadsest ja raskesti ümber sõnastatavast tähendusest ja
- kolmandaks probleem, et *wordnet*-tüüpi tesaurused ei käsitle hetkel eessõnu veel iseseisva sõnaliigina.

Peale selle on paljud püsiühendid oma suure varieerumise tõttu üldse sobimatud, et neid *wordnet*-tüüpi tesaurusesse lisada.

Seoses liitsõnadega on aga probleemiks see, et eesti keeles ei ole liitsõnade hulk piiratud ning iga keeletarvitaja võib neid vastavalt vajadusele ise luua. Leksikoni üldiselt ja ka konkreetselt *wordnet*-tüüpi tesaurustesse ei saa ega ole mõtetki neid lõputult lisada.

Et saada ülevaade, kuidas on püsiühendid ja liitsõnad esindatud Eesti Wordnet'is uurisin sagedasemaid verbikeskseid püsiühendeid ja levinumatest nimisõnadest moodustatud liitsõnu. Kuigi enamik ligi 40 uuritavast püsiühendist oli tesaurusesse juba lisatud, esines ebatäpsusi nende semantilistes suhetes ning puudu olid mitmed laialt

kasutatavad sünonüümid. Töö liitsõnade lisamise suhtes tundus aga võrreldes püsiühenditega olevat puudulikum. Eesti Wordnet'is ei ole mitmeid keeles sageli kasutatavaid tavalisi liitsõnu ning puudu on ka palju metafoorsed liitsõnu, mis on aga eriti just eesti kõnekeeles väga laialt levinud.

Niisiis võib kokkuvõtvalt öelda, et nii püsiühendite kui ka liitsõnade lisamine peaks saama tulevikus Eesti Wordnet'i tegijatele üheks prioriteediks. Üldiselt võiks just liitsõnade, aga ka püsiühendite, lisamisel arvestada nende sageduse ja tähenduslikkusega keeles. Selleks võiks näiteks tekstikorpusest välja otsida sagedasemad püsiühendid (lisaks verbikesksetele püsiühenditele ka muud) ja liitsõnad ning seejärel kontrollida, kas tegemist on piisavalt keeles kasutatud ja iseseisvate täistähenduslike mõistetega ning juhul kui on, need siis ka tesaurusesse lisada.

Idiomaatiliste püsiühendite lisamiseks võiks kaaluda inglisekeelse WordNet'i tegijate poolt välja pakutud metafoorse ehk idiomaatilise suhte loomist, mis aitaks vältida idiomaatiliste ühendite tähenduse ähmastumist nende lihtsalt lisamisel üldisematesse sünohulkadesse. Samuti idiomaatiliste ühendite tähenduste osadeks jaotamist ja nende üksikult suhestamist teiste sünohulkadega. Ka idiomaatiliste liitsõnade tähenduse aitaksid sellised lahendused paremini välja tuua.

Kirjandus

Baran, *Anneli* 2004. Fraseoloogilistest liitnimisõnadest. Mäetagused, nr 27. Tartu : Eesti Kirjandusmuuseum, lk 157-166.

Carter, *Ronald*; *McCarthy, John* 2006. Cambridge Grammar of English. Cambridge: Cambridge University Press.

Eesti Wordnet; http://www.cl.ut.ee/ressursid/teksaurus

Eesti keele seletav sõnaraamat; http://www.eki.ee/dict/ekss/

Eesti õigekeelsussõnaraamat; http://www.eki.ee/dict/qs2006/

EKG I 1995 = *Erelt, Mati*, *Kasik*, *Reet; Metslang, Helle; Rajandi*, *Henno; Ross*, *Kristiina; Saari*, *Henn; Tael*, *Kaja; Vare, Silvi* 1995. Eesti keele grammatika I. Morfoloogia. Sõnamoodustus. Tallinn: Eesti Teaduste Akadeemia Eesti Keele Instituut.

EKK 2000 = *Erelt*, *Mati*; *Erelt*, *Tiiu; Ross, Kristiina* 2000. Eesti keele käsiraamat. Eesti Keele Sihtasutus. Tallinn.

Fellbaum, Christiane 1998. Towards a representation of idioms in WordNet. In S. Harabagiu (ed), Proceeding of the Workshop on Usage of WordNet in Natural Language Processing Systems. Montreal: COLING/ACL, pp. 52-57.

Fellbaum, Chistiane (ed) 2006. Corpus-Based Studies of German Idioms and Light Verbs. Special issue of the Journal of Lexicography, 19.4.

Kaalep, *Heiki-Jaan*; *Muischnek*, *Kadri* 2002. Eesti kirjakeele sagedussõnastik. Tartu: Tartu Ülikooli Kirjastus.

Krikmann, *Arvo* 2004. „Sai hea obaduse vastu obadust“: Löömist ja peksmist märkivad väljendid eesti keeles. Reetor 3. Tartu: Eesti Kirjandusmuuseum, Folkloristika osakond, Eesti Kultuuriloo ja Folkloristika Keskus, lk 52–144.

Langemets, *Margit* 2004. Polüseemia ja leksikograafia. Emakeele Seltsi aastaraamat. Toim. Mati Erelt, Maria-Maren Sepper. Tallinn: Emakeele Selts, lk 97-124.

Lönneker, *Birte*; *Carina, Eilts* 2004. A Current Resource and Future Perspectives for Enriching WordNets with Metaphor Information. 2nd International Conference of the Global WordNet Association, Brno, Czech Republic (GWC-2004).

Muischnek, *Kadri* 2006. Verbi ja noomeni püsiühendid eesti keeles. Dissertationes Philologiae Estonicae Universitatis Tartuensis 17. Tartu: Tartu Ülikooli Kirjastus.

Nunberg, *Geoffrey*; *Thomas Wasow*; *Ivan A. Sag* 1994. Idioms. Language. Vol. 70, No. 3. pp. 491-538.

Orav, Heili; *Vider, Kadri* 1998. Sõna tasandilt mõiste ruumi. Keel ja Kirjandus, nr 1, lk 57- 64.

Orav, Heili; *Vider, Kadri* 2006. Millist leksikoni vajab arvuti tähenduse mõistmiseks? – Keel ja Arvuti. Toim. Mare Koit; Renate Pajusalu; Haldur Õim. Keel ja arvuti. Tartu: Tartu Ülikooli Kirjastus, lk 85-96.

Osherson, *Anne; Fellbaum*, *Christiane* 2010. The Representation of Idioms in WordNet. International Conference of the Global WordNet Association, Mumbai (GWC-2010).

Sag, Ivan A.; *Baldwi, Timothy; Bond, Francis; Copestake, Ann; Flickinger, Dan* 2001. Multiword Expressions: A Pain in the Neck for NLP. LinGO Working Paper No. 2001-03.

Sharada, B. A.; Girish, P. M. 2004. WordNet Has No `Recycle Bin'. Proceedings of the Second Global WordNet Conference, Brno, Czech Republic, pp. 311-319.

Sünonüümisõnastik; http://dict.ibs.ee/syn/

WordNet; http://wordnet.princeton.edu/

Õim, Asta 1993. Fraseoloogiasõnaraamat. Tallinn: Eesti Teaduste Akadeemia Keele ja Kirjanduse Instituut.

Summary

The aim of the present work is to investigate frequent linguistic phenomena such as multi-word expressions and compound words in wordnet-type thesauri, concentrating on Estonian Wordnet.

A multi-word unit is a combination of two or more words that occur together to express a single meaning. In English, compound words are often written separately and therefore seen as a kind of multi-word expression. In Estonian, compounds are almost always written as single words and therefore separated from multi-word expressions. The fact that there is no certain definition for neither of these expressions makes it also difficult to include them in wordnets. There are several problems that occur when adding them, for example formal and semantic problems as well as some more specific problems like handling prepositions in the wordnet structure. Besides, some idiomatic constructions are just too complex and variable to integrate them.

Including compound words into wordnet-type thesaurus is a problem for Estonian language as well as for example for the German language, because in both of these languages words can be combined quite freely while the meaning still stays understandable. Nevertheless, the number of compounds in wordnets should be somehow restricted.

As a result of an analysis of the most frequent verbal multi-word expressions and compound nouns in Estonian Wordnet, it can be said that although there are many multi-word expressions already included, inaccuracies in semantic relations and missing synonyms are rather frequent.

The number of compounds in Estonian Wordnet seems to be small compared to their frequent use in the Estonian language. Some very common compounds, especially figurative compounds are not yet included in Estonian Wordnet.

To sum up, the editors of Estonian Wordnet should definetely include more multi-word expressions and compound words in Estonian Wordnet and thereby take into account the characteristics of these linguistic phenomena.