doi:10.5128/ERYa8.09

EESTI RAKENDUSLINGVISTIKA ÜHINGU AASTARAAMAT 8, 139–151

EESTI WORDNET'I STRUKTUURI ANALÜÜSIST

Ahti Lohk, Leo Võhandu

Ülevaade. Artikkel pakub üldise lähenemisviisi relatsiooniliste süsteemide suletud hulkade leidmiseks ja korrastamiseks ning demonstreerib selle meetodi kasutamist Eesti Wordnet'il. Lahatakse Eesti Wordnet'i struktuuri, lähtudes semantilistest suhetest. Selgitatakse analüüsiks kasutatava infotöötlusmeetodi ideed ja sõnastatakse lahendatav probleem mitteformaalselt. Esitatakse meetodi rakendamise järjestikused sammud. Andmetöötluse tulemusena tekivad visuaalsed analüüsi objektid/pildid, mis avavad Wordnet'i struktuuri viisil, mis võimaldab leksikograafil hinnata struktuurides peituvaid eripärasid. Artikli lõpuosas antakse näidete põhjal vihjeid võimalikele probleemidele ja nende lahendustele.

Võtmesõnad: tesaurus, semantilised suhted, Wordnet'i visualiseerimine, suhete järjestamine, eesti keel

1. Üldist

Kevadel 2011 toimunud rakenduslingvistika konverentsil autorite esitatud ettekanne "Eesti sünovara visualiseerimisest" tegi sisekaemuse kahte, Internetist kättesaadavasse, eesti keele mõistelisse sõnaraamatusse ehk tesaurusse: Filosoft OÜ koostatud tesaurus[1] ja Tartu Ülikooli eesti keele tesaurus (TEKSaurus) ehk Eesti Wordnet[2]. Eesmärgiks oli uurida sõnastike sisestruktuure ja esitada neid tabelite, jooniste ning graafikute kaudu. Samuti pakuti välja võimalusi Eesti Wordnet'i andmete struktuurseks esitamiseks (Lohk, Võhandu 2011).

Käesolevas artiklis piirdume siiski vaid ühe sõnastiku, Eesti Wordnet'iga. Kohe alguses tasuks mainida, et Eesti Wordnet'i loomisel on võetud eeskujuks nii mitmekeelne EuroWordNet[3] kui ka algne ingliskeelne, Princetoni ülikoolis loodud Wordnet[4], mis oma olemusliku struktuuri poolest on eeskujuks eri keelte *wordnet*-tüüpi sõnastikele (Orav jt 2011). Selliste sõnastike loomine ja arendamine on üleilmsete mõõtmetega ettevõtmine, millest annavad tunnistust

[1] Filosoft OÜ tesaurus vt http://www.filosoft.ee/thes_et/ (1.10.2011).
[2] Tartu Ülikooli TEKSaurus ehk Eesti Wordnet vt http://www.cl.ut.ee/ressursid/teksaurus/teksaurus.cgi.et/ (1.10.2011).
[3] EuroWordNet vt http://www.illc.uva.nl/EuroWordNet/ (1.10.2011).
[4] Wordnet vt http://wordnet.princeton.edu/ (1.10.2011).

iga kahe aasta tagant toimuvad konverentsid pealkirjaga "The Global Wordnet Conference"[5].

Nii Eesti Wordnet kui teised *wordnet*-tüüpi tesaurused omavad sõnade organiseeritust sünohulkadesse, kuhu on koondatud ühte mõistet väljendavad sünonüümsed sõnad ja sõnaühendid. Sünohulgad võivad olla ka üheliikmelised, sest mõistet võib esindada keeles ka ainult üks sõna.[6] Seega tingib sünohulga tesaurusesse lisamise mõiste eksisteerimine mentaalses leksikonis, mitte sünonüümiasuhe mõistet esindavate sõnade vahel keeles.

Sünohulkades olevate sõnade eristamiseks on igale sõnale lisatud tähendusindeks ja sõnaliik. Kolmik {*sõna_tähendusindeks_sõnaliik*} annab sõnale n-ö tähendusliku unikaalsuse. Kõigis illustreerivates joonistes arvestatakse just selle kolmest osast koosneva tervikuga.

Sünohulgad on omavahel seotud semantiliste suhete kaudu, mida Eesti Wordnet'is on 43 erinevat liiki, nt hüperonüümiaseos, hüponüümiaseos, antonüümiaseos jne.

Eesti Wordnet'i süvastruktuurne uurimine tõi esile mõningaid ebakõlasid. See on ka loomulik, sest *wordnet*-tüüpi tesauruse puhul on tegu vägagi keerulise ja paljurelatsioonilise süsteemiga.[7]

Esialgsed tulemused näitasid kätte tee *wordnet*-tüüpi süsteemide kiireks ja usutavasti leksikograafile kasulikuks analüüsiks. Tegemist on infoteoreetilises mõttes keeruka probleemiga, mille jaoks ei ole siiani olemas täielikku formaliseeritud lahendusmeetodit. Graafiteooria meetodeid kombineerides õnnestus luua programmipakett, mis teeb *wordnet*'i mahuga süsteemide süvaanalüüsi mõistliku ajaga ära. Tulemuste esitluseks kasutatakse sellist visualiseerimismeetodit, mis usutavasti annab leksikograafile kergesti käsitletava vahendi *wordnet*-tüüpi sõnastike hetkeolukorda kajastava struktuuri uurimiseks, hindamiseks ja ebakõlade kõrvaldamiseks.

Vahemärkusena tuleb öelda, et ka Eesti Wordnet'i sisestusliides Polaris võimaldab mõnesugust visualiseerimist, kuid jälgib seejuures leksikograafi valitud rada, tegemata süsteemi sõltumatut üldkontrolli, mis on selle artikli peaeesmärk.

Järgnevas esitatakse mitteformaalne ülevaade kasutatud matemaatilisest metoodikast, demonstreeritakse paari sõnastikus olevate erirelatsioonide kinniste alamhulkade eraldamise tüüpilist tulemust ja näidatakse, kuidas neid tulemusi edasiseks tõlgendada.

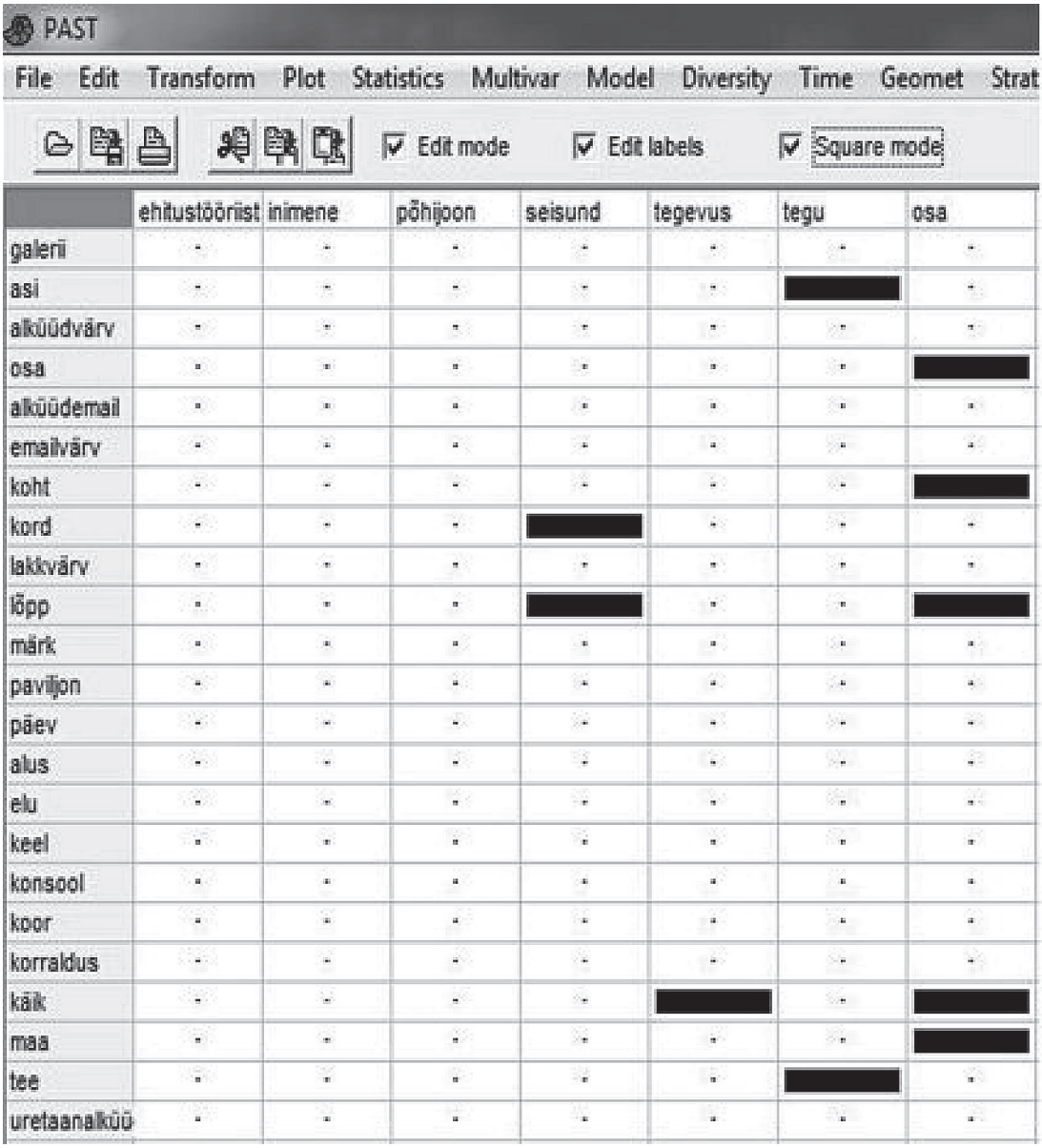
Artikli piiratud mahu tõttu piirdume ainult nimisõnade hüperonüümse seosega määratud allsüsteemi uurimisega. Eraldasime Eesti Wordnet'ist suure (40099 x 5887) tabeli, mille olemusest annab aimu allpool esitatud väike väljavõte (vt tabel 1).

[5] The Global WordNet Association vt http://www.globalwordnet.org/ (1.10.2011).

[6] Üheliikmelised sünohulgad moodustavad tesauruses hetkel 57% kõigist sünohulkadest ning kõige rohkem on ühes sünohulgas 20 liiget (Orav jt 2011: 97).

[7] Muu hulgas selgus näiteks, et autorid olid heauskselt kasutanud Eesti Wordnet'i veebivarianti, mis osutus viis versiooni värskeimast vanemaks. Tänu Neeme Kahuski lahkele vastutulekule õnnestus peagi saada sõnastiku viimane versioon (versioon 60, 25.04.2011) Eesti Wordnet'i arendusvahendina kasutatava sisestusliidese Polaris exportfail txt-formaadis. Edasine koostöö tesauruse kollektiiviga ongi kulgenud õige tulemusrikkalt.

Tabel 1. Suhtetabel, milles hüponüüm-sünohulkade esimesed liikmed (sõnad) on ridades ja hüperonüüm-sünohulkade liikmed (sõnad) on veergudes



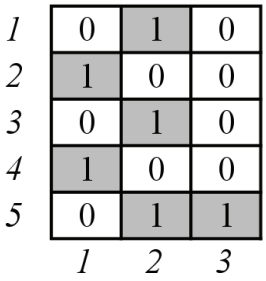
	ehitustööriist	inimene	põhijoon	seisund	tegevus	tegu	osa
galerii	·	·	·	·	·	·	·
asi	·	·	·	·	·	■	·
alküüdvärv	·	·	·	·	·	·	·
osa	·	·	·	·	·	·	■
alküüdemail	·	·	·	·	·	·	·
emailvärv	·	·	·	·	·	·	·
koht	·	·	·	·	·	·	■
kord	·	·	·	■	·	·	·
lakkvärv	·	·	·	·	·	·	·
lõpp	·	·	·	■	·	·	■
märk	·	·	·	·	·	·	·
paviljon	·	·	·	·	·	·	·
päev	·	·	·	·	·	·	·
alus	·	·	·	·	·	·	·
elu	·	·	·	·	·	·	·
keel	·	·	·	·	·	·	·
konsool	·	·	·	·	·	·	·
koor	·	·	·	·	·	·	·
korraldus	·	·	·	·	·	·	·
käik	·	·	·	·	■	·	■
maa	·	·	·	·	·	·	■
tee	·	·	·	·	·	■	·
uretaanalküü	·	·	·	·	·	·	·

2. Infotöötlusmeetodi idee selgitus

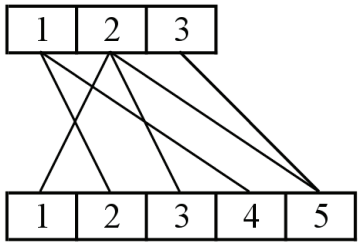
Selgitame kasutatava meetodi põhiolemust lihtsa kunstliku näitega.

Olgu antud objektide ja neid kirjeldavate tunnuste suhtemaatriks. Selle maatriksi ridadeks olgu Eesti Wordnet'i hüponüüm-sünohulgad ja veergudeks hüperonüüm-sünohulgad, s.t me kasutame selles näites vaid ühte suhet tesauruse paljudest suhetest.

Näitemaatriks on joonisel 1a (1 näitab mingi hüponüüm-sünohulga seotust teise, hüperonüüm-sünohulgaga).



Joonis 1a. Suhtemaatriks



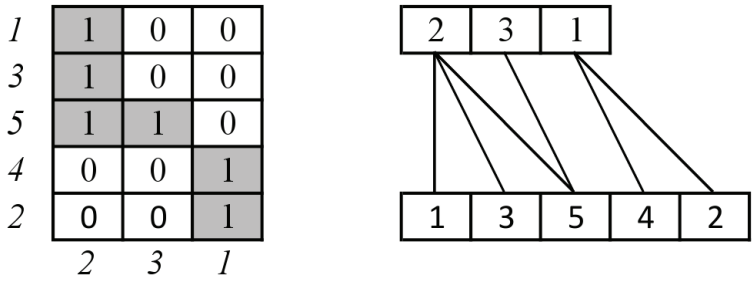
Joonis 1b. Kahealuseline graaf

Seda maatriksit saab esitada kahealuselise graafina, kus ühel joonel on objektide (ridade) järjenumbrid ja teisel joonel on tunnuste järjenumbrid (vt joonis 1b). Joonisel 1b on ülemisel joonel veerunumbrid ja alumisel joonel ridade numbrid.

Jooniselt 1b on kerge loendada, et joonte ristumisi on parajasti kolm. Kui objekte ja tunnuseid on palju (kümned, sajad, tuhanded, sajad tuhanded jne), siis muutuks sellise joonise koostamine peaaegu võimatuks.

Püstitame endale küsimuse: kas tabeli ridu ja veerge ümber järjestades ei oleks võimalik ristumiste arvu vähendada? Selline ekstreemumprintsiipidest lähtuv mõtteviis on matemaatikule omane.

Praeguse triviaalnäite korral on kerge otsese kontrolliga veenduda, et muutes ülemisel joonel tunnuste järjestuse (123) asemele (231) ja alumisel joonel ridade (12345) asemele (13542), saame korrastatud suhtemaatriksi ja vastava graafi (vt joonis 1c).



Joonis 1c. Korrastatud suhtemaatriks ja kahealuseline graaf

Joonise 1c parempoolse osa põhjal on kohe näha, et esiteks ei ole enam ühtegi ristumist ja teiseks lagunes algne joonis 1b kaheks seotud tükiks.

Need lihtsad faktid on tähtsad, sest nad näitavad, et algne andmemaatriks laguneb kaheks sidusaks alamhulgaks (kinniseks hulgaks). Teiseks tekkis tunnuste ja objektide jaoks loomulik lineaarne järjestus, millele tavaliselt praktikas vastab samuti mingi loomulik semantiliselt sisuline järjestus.

Nüüd on kerge ka algmaatriks ümber järjestada uude järjestusse (vt joonis 1c vasakpoolne osa). Selline ribamaatriks on psühholoogias tuntud multidimensionaalse järjestusskaalana (MDSCALE). Praegu me selle tulemuse juures ei peatu, aga Eesti Wordnet'i sisestruktuuri uurimisel võib sellest kasu olla.

3. Infotöötlusprobleemi mitteformaalne sõnastus

Järgnevalt püstitame endale ülesande leida sobiv meetod ja koostada algoritm ning programm, mille abil saada algandmete tabeli kahealuselisest esitusest minimaalse võimaliku joonte ristumiste arvuga esitus (reaalsetes andmetes ei pääse me lõpptulemuses ristumistest!). See probleem on graafiteoorias tuntud nime all "kahealuselise graafi ristumiste arvu minimeerimine" ning see on tõestatult NP-keerukas probleem.[8] See tähendab, et lõplikku polünoomiaga töötavat algoritmi, mis garanteeriks vajaliku miinimumi, pole olemas (Garey, Johnson 1979). Mida siis teha?

Esimese vahesammuna – veel enne, kui asuda ristumiste minimeerimise juurde – saab lihtsa võttega andmetabelist eraldada kõik seotud kinnised hulgad. Nii suure andmekogu juures nagu Wordnet, on see väga suur ajavõit ja kergendab ühtlasi ka saadud tulemuste hilisemat tõlgendamist.

Meie pisinäites kasutatud maatriksesitus (vt joonis 1c) võtab suurte tabelite korral liiga palju mälu ja teeb ka algoritmid aeglaseks. Seepärast kasutasime Eesti Wordnet'i andmete jaoks nimistulist esitust (ehk *list*-esitust), kus iga sõna jaoks antakse nimekiri, millised sõnad on algsõnaga vastavas suhtes, hoides märkimisväärselt mälumahtu kokku. Selline esitus on tingitud Polarise eksportfaili andmetest, kus ükski sünohulk ei ole otseselt seotud ühegi teise sünohulga identifitseeriva numbri ega sõnade loendiga, vaid on viidud vastavusse (on semantilises seoses) teise sünohulga esimese liikmega (sõnaga).

Teine väga oluline võte seisnes selles, et teisendasime kahealuselise graafi lihtsal viisil nn intervallgraafiks, viies ülemisel joonel olevad elemendid alumisele joonele seal juba paiknevate elementide järele. Loomulikult tuleb muuta nende elementide järjenumbreid. Liidame lihtsalt igale ülemise joone elemendi numbrile juurde arvu N (s.o objektide arvu). Selle teisenduse järel on tegu standardse intervallgraafiga (seosed jäävad muidugi alles!). Knuthi (1968) programmeerimisentsüklopeedias ja arvukate algoritmide raamatus (Flannery jt 2009: 346) leidub mõnus algoritm, mis võimaldab eraldada kõik algse andmetabeli kinnised hulgad väga kiiresti. Eesti Wordnet'i kõigi kinniste hulkade eraldamine hüperonüümsete suhete korral võttis (tavalisel sülearvutil) aega ainult 2,5 minutit.

Alles kõigi kinniste hulkade eraldamise järel kasutame Niermanni (2005: 41–46) kirjeldatud evolutsioonilise optimeerimise ideed, et ristumiste arv viia igas eraldatud kinnises hulgas miinimumini.

Selline samm-sammuline lähenemine on osutunud väga efektiivseks suurte kahendmaatriksite sisemise struktuuri uurimisel. Muidugi tuleb paraku märkida, et infotöötlusliku käsitluse täielik esitus jääb vastavasisulise ajakirja teemaks.

4. Näiteid eraldatud kinnistest hulkadest

Toodud valik eraldatud kinnistest hulkadest (vt tabel 2; joonis 2, 3, 4, 5) on illustratiivse iseloomuga, et anda aimu meetodiga saadavatest tulemustest. Tabeli 2 igal real on järgmine sisu: taustatoonita lahtrites on hüponüüm-sünohulkadest pärinevad sõnad ja tooniga lahtrites hüperonüüm-sünohulki esindavad sõnad, millel on semantiline seos hüponüümidega. Et ideaaljuhul on igal sünohulgal vaid üks hüperonüüm (Vider 2001), annavad hüperonüüm-sünohulka esindavad

[8] NP (ingl *non-deterministic polynomial*) – mittedeterministlikult polünomiaalne.

sõnad[9] leksikograafile kiire ülevaate, kui paljude kõrgema taseme sünohulkadega on hüponüümid seotud. Nii on nt tabeli 2 esimesel real sünohulka esindavateks sõnadeks "hotell" ja "ujuvrajatis". See tähendab, et kogu real kui kinnisel hulgal on kaks hüperonüüm-sünohulka. Millisest või millistest sünohulkadest pärinevad "flotell", "pontoonkai" ja "ujuvkai", ei ole siin tähtis teada. Kõikidel valge taustaga ridadel on ühe või mitme sünohulga sõnad täies mahus.

Tegelikus töös koostab programm KLASSID kõigepealt vastavalt uurija soovile suurest Eesti Wordnet'i andmefailist töötlusse mineva suhtetabeli (vt tabel 1). Järgmise sammuna koostatakse kinniste sidusate alamhulkade loetelu. Tükike sellest tabelist on nähtav tabelis 2.

Tabel 2. Fragment kinniste alamhulkade tabelist

flotell_1_n	pontoonkai_1_n	ujuvkai_1_n	hotell_1_n	ujuvrajatis_1_n
frakk_1_n	vatt_3_n	kuub_1_n	peigmehe rõivastus_1_n	ülikond_1_n
grammofon_1_n	plaadimasin_1_n	vinüülplaadi-mängija_1_n	fonograaf_1_n	taas-esitusseade_1_n
kabuhärg_1_n	kabuvärss_1_n	kohihärg_1_n	härg_1_n	kohiloom_1_n
muusikahall_1_n	black box_1_n	hall_2_n	teatrisaal_1_n	

Kõige suurem hüperonüümiasuhte abil Eesti Wordnet'ist eraldatud sidus kinnine alamhulk on mõõtmetega 4945 x 457 (hüperonüüm-sünohulka esindavat sõna). Kinniseid alamhulki tekkis üldse 6051. Jaotumine vastavalt: 4854 nimisõna, 1178 tegusõna ja 19 omadussõna hulka.

Kõigi semantiliste suhete ja liikide samaaegsel kasutamisel tekib kinniseid alamhulki *ca* 500.

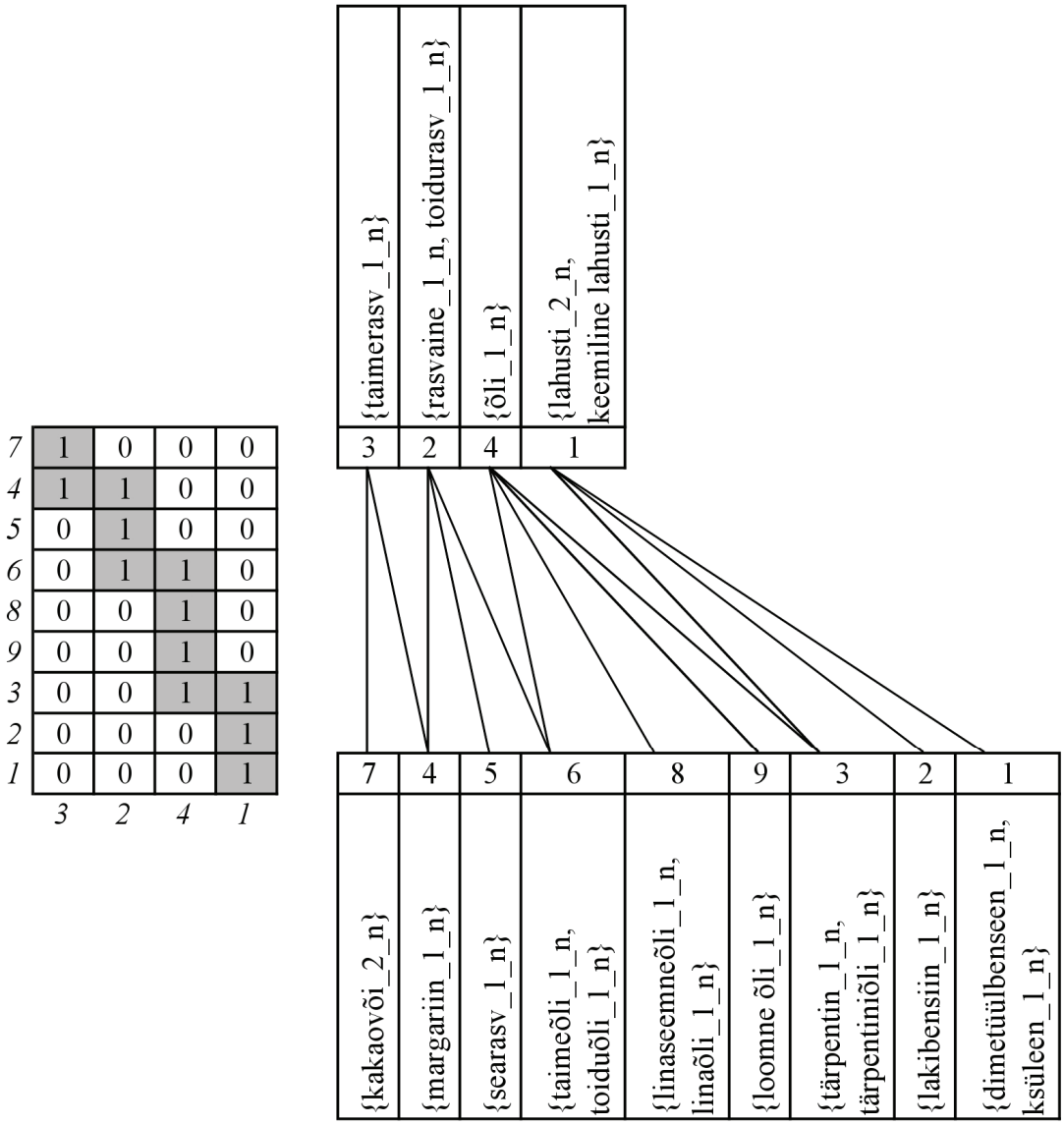
Kolmanda sammuna valime nendest ribadest (vt tabel 2) suvalise meid huvitava seotud hulga ja laseme programmil MINCROSS luua minimaalsete ristumiste arvuga pildi. Tüüpiline väljatrükk tehtud tööst on niisugune, kus nii vasakpoolse ümberjärjestatud tabeli kui ka parempoolse joonise äärtel olevad numbrid osutavad sellele analüüsile eelnenud riba väljatrüki järjenumbritele (vt joonised 2, 3, 4, 5). Sellised minimaalsete ristumiste arvuga joonised on aluseks tekkinud kinniste hulkade edasisele sisulisele analüüsile.

Wordnet-tüüpi sõnastikes on semantilised seosed esitatud sünohulkade vahel. Kuigi tutvustatava meetodi puhul ei ole tähtis, millised objektid millistes seostes esinevad, on antud artiklis jäädud siiski vaate "seosed sünohulkade vahel" juurde (vt joonised 2, 3, 4, 5).

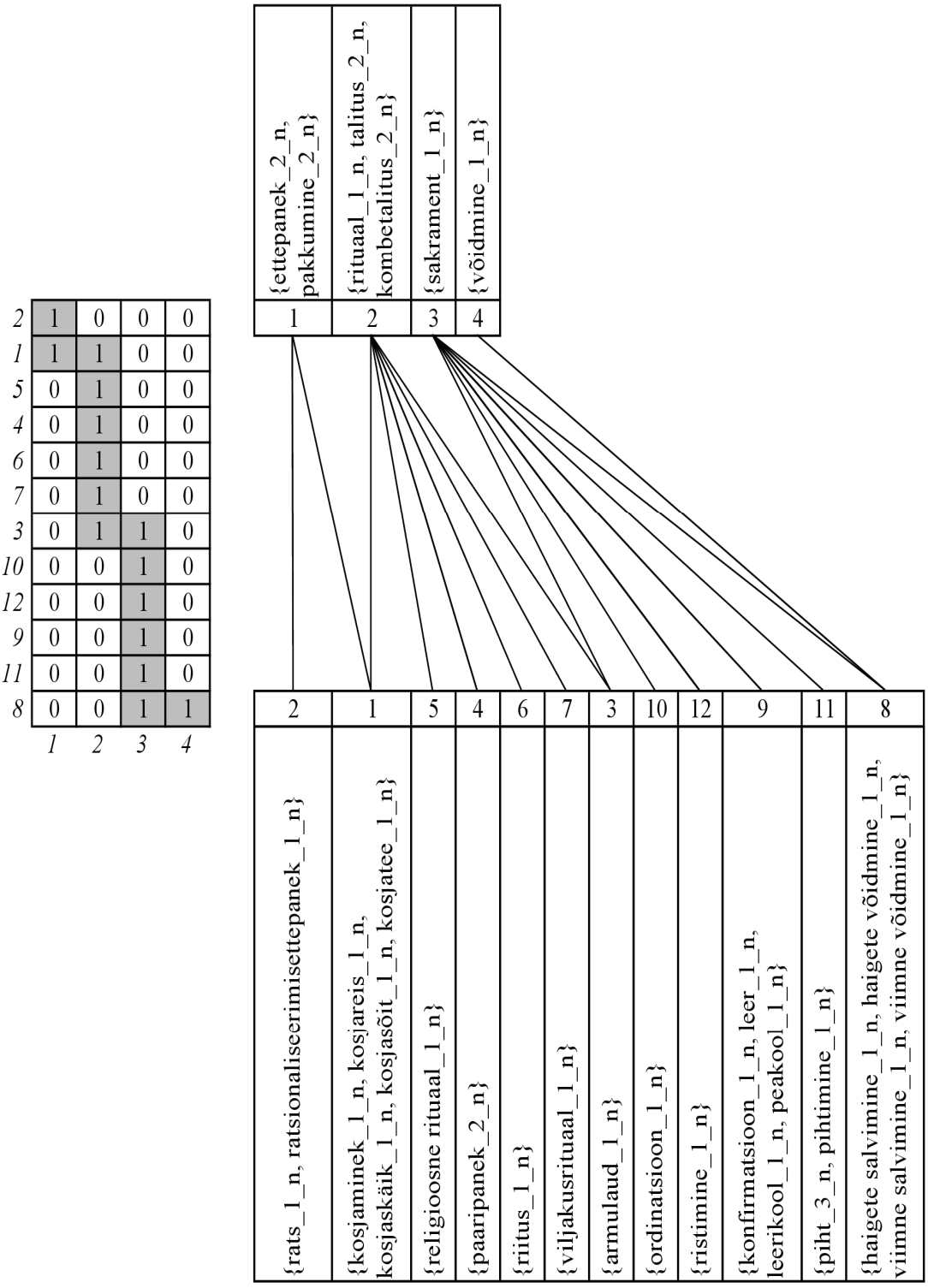
Pisipilt joonisel 4 on hea näide (leksikograafi) mingil põhjusel pooleli jäänud töö kohta. Pildilt puudub rida teisi spordialasid või vähemalt hüppeid.

Niisuguse pildi nägemine peaks leksikograafi ergutama muude spordiga seotud sõnade ülesotsimisele, nende seospiltide loomisele ja siis ühendklastri loomisele. Nii võib Eesti Wordnet'i lähemal uurimisel leida, et "kolmikhüpe" on sidumata kergejõustikuga, samas kui teised samalaadsed kergejõustikualad (teivashüpe, kaugushüpe, kõrgushüpe jt) on sellega seotud. Probleemi olemasolu kinnitab ka kergejõustiku kui sünohulga definitsioon: "üldisem spordiala, mille alla kuuluvad jooksud, kiirkäimine, hüpped, heited ja mitmevõistlus".

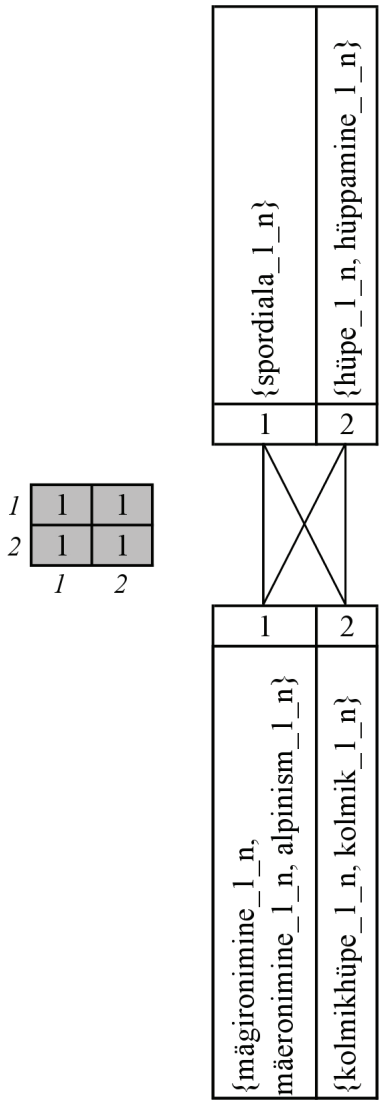
[9] Eesti Wordnet'i veebirakenduses kasutatav formaat esitab samuti sünohulga hüperonüümiks vaid ühe (esimese) hüperonüüm-sünohulka kuuluva sõna. Sealt on artikli autorid ka eeskuju saanud.



Joonis 2. Näide (1), ümberjärjestatud suhtemaatriks ja minimaalse ristumiste arvuga kahealuseline graaf



Joonis 3. Näide (2), ümberjärjestatud suhtemaatriks ja minimaalse ristumiste arvuga kahealuseline graaf



Joonis 4. Näide (3), ümberjärjestatud suhtemaatriks ja minimaalse ristumiste arvuga kahealuseline graaf

5. Vihjeid saadud piltide edasiseks analüüsimiseks

Vaatleme näitena pilti, kus peaks huvi äratama hüperonüümide "võidmine" ja "sakrament" vahekord (vt joonis 3). Programmi MINCROSS töö tulemusel saadud pildilt on näha, et hüperonüümi "võidmine" kõik seosed kuuluvad "sakramendi" alla. Järelikult on oodata, et "sakrament" oleks "võidmise" hüperonüüm. Edasine ebakoht seoste skeemis on seotud sõnaga "armulaud". Põhimõtteliselt õigesti on "armulaud" seotud nii "sakramendi" kui "rituaaliga". Ometi on ju kõik "sakramendi" hüponüümid "talitused" (toimingud). Need seosed on ilmselt süsteemist puudu.

Lisaks veel üks pisipeensus. Kuigi abielu kuulub sakramendi hulka, on Eesti Wordnet'is ta liigitatud vaid rituaaliks ja mitte otseselt, vaid kaudselt sõna "paaripaneku" kaudu!

Joonisel 2 on sattunud kokku "lakibensiin" ja "margariin". Esimene neist on lahusti, teine aga Eesti Wordnet'i andmetel lahustite hulka ei kuulu. Margariini ja lakibensiini seostatus oleks ilmselt põhjendatud, kui esimene neist omaks (semantilist) seotust toiduainekeemiaga.

Äsjane informaatikute tehtud lühianalüüs on ilmselt üpris ebatäielik, aga ometi peaks see osutama MINCROSS programmi pakutavatele huvitavatele võimalustele Eesti Wordnet'i sisestruktuuri üksikasjalikuks süvauurimiseks.

Lõpuks veel üks lisavõimalus, mille pakub juursünohulga kasutusele võtmine.

Juursünohulk on siin defineeritud kui sünohulk, mis rahuldab samaaegselt kahte tingimust: 1) on vähemalt ühele sünohulgale hüperonüüm-sünohulgaks ja 2) talle endale ei ole vastavusse seatud ühtegi hüperonüüm-sünohulka. Ehk sünohulk, millel on alluvad, kuid puuduvad ülemused. Neile tingimustele vastavaid sünohulki on Eesti Wordnet'is 185. Jagunemine vastavalt: 142 nimisõna, 24 tegusõna ja 19 omadussõna sünohulka.

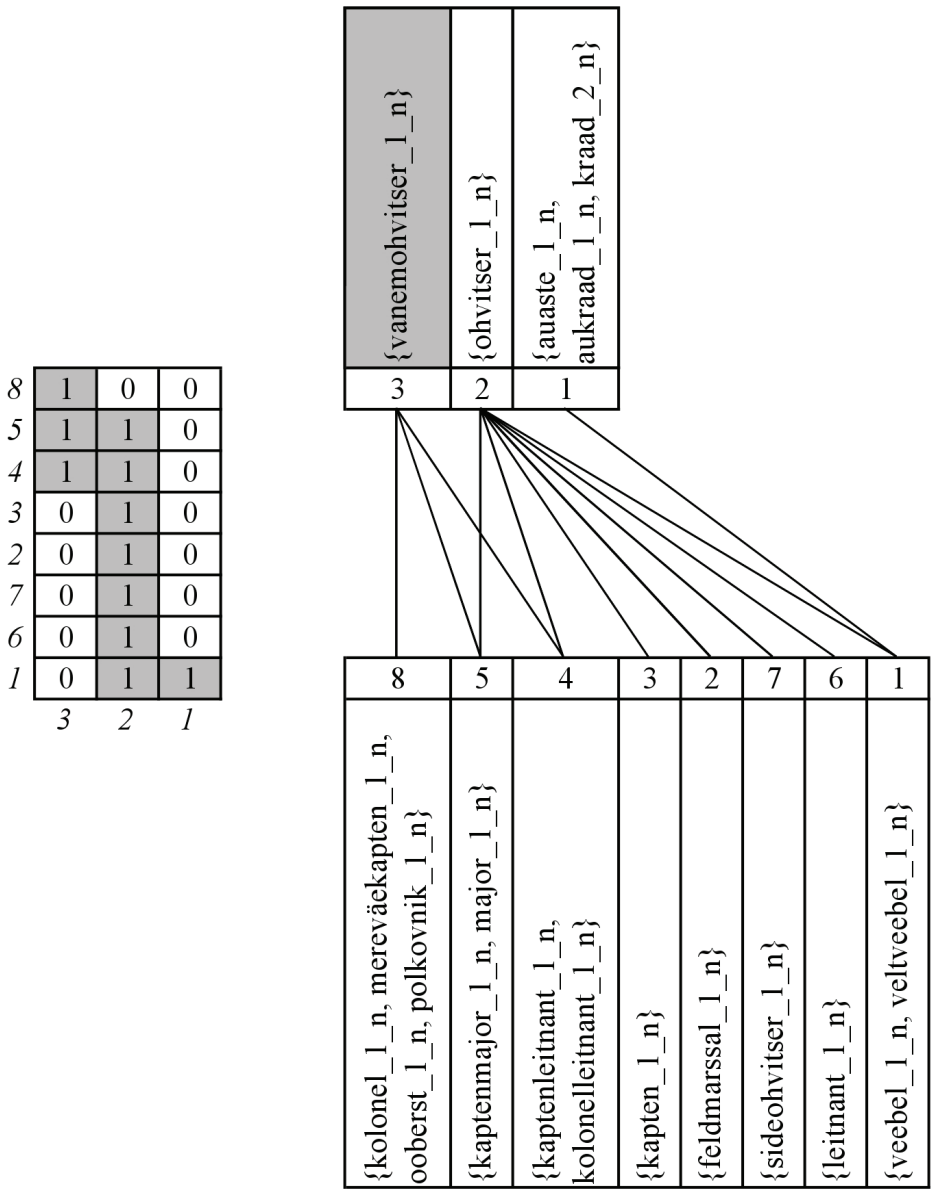
Juursünohulkade läbivaatamine oleks muidugi omaette teema. Meid aga huvitab kahealuseliste graafide puhul see, kas esineb selliseid graafe, mille hüperonüümiosa (kahendgraafi ülemine rida) sisaldab muuhulgas juursünohulki. Kinnise hulga hüperonüümsünohulgale saab tuua juurde uued hüperonüümiaseosed. Niimoodi järjest üles liikudes mõistame selgesti, et sünohulgad paiknevad oma hüperonüümiapuus eri tasemetel. Juursünohulkade esinemine kahealuselise graafi hüperonüümiosa osas võiks sellisel juhul viidata kas võimalikule probleemile eri taseme sünohulkade kasutamise kohta, mis ei pruugi küll alati probleemiks osutuda, kuid mis võib juhatada pooleli jäänud (semantilisele) seostamisele. Sellist näidet illustreerib joonis 5: {*vanemohvitser_1_n*} on siin juursünohulk (eristatud halli taustaga). S.t sellel sünohulgal endal ei ole ühtegi hüperonüüm-sünohulka. Täpsemal uurimisel selgubki, et ohvitseri ülemmõisteks on sõjaväelane, mis on ju ka "vanemohvitser". Antud näite puhul hakkab veel silma, et {*auaste_1_n*, *aukraad_1_n*, *kraad_2_n*} on sidumata auastmeid sisaldavate sünohulkadega.

Lõpetuseks, kui vaadata näiteks hüperonüüm-sünohulki (vt joonis 3) {*ettepanek_2_n*, *pakkumine_2_n*}, {*rituaal_1_n*, *talitus_2_n*, *kombetalitus_2_n*}, {*sakrament_1_n*}, {*võidmine_1_n*}, siis pärast seda, kui leksikograaf on sõltuvused korrastanud, võib see korrastus aidata kergemini käsitleda taksonoomilist lisatükeldust paljude alluvatega ülemmõistete korral (Orav jt 2011: 101, Kozareva, Hovy 2010: 1110-1111)[10]. Näiteks joonisel 2 esitatud ülemmõisted "taimerasv" ja "lahusti" viitavad selgele vajadusele täpsustada taksonoomilist struktuuri. Pildil eksisteeriv seoste süsteem ise annab aga võimaluse luua korralik side ka hüperonüümide vahel ning saada kõrgema taseme seosgraaf, mis n-ö ronib antud relatsiooni abil juurmõistele (juursünohulgale) lähemale. Sellise automaatse või interaktiivse pealisehituse loomine on muidugi omaette tõsine töö.

6. Kokkuvõte

Elu on näidanud, et sageli saavutatakse mingis valdkonnas edu, kui sekkuvad "mitteeksperdid". Seetõttu esitavadki autorid – informaatikud – Eesti Wordnet'i jaoks spetsiifilisi andmetöötluse ja visualiseerimise meetodeid. Piltide ja tabelitena esitatud tagasisidemehhanismid sobivad iga *wordnet*-stiilis loodud süsteemi jaoks, võimaldades semantilist analüüsi ja seaduspärasuste uurimist. Leksikograaf saab kasutada uudset lähenemist, et uurida süsteemi varjatud struktuure ja teha vajalikke parandusi.

[10] Hoolimata paljudest katsetest ei ole täisautomaatset taksonoomilist tükeldust/liigitamist veel teostatud, *wordnet*-tüüpi sõnastikes koostatakse need aga käsitsi. Nii on loomulik, et tekivad korrastust vajavad vead.



Joonis 5. Näide (4) – juursünohulk, ümberjärjestatud suhtemaatriks ja minimaalse ristumiste arvuga kahealuseline graaf

Artiklis kasutatud mitmed keerukad andmeanalüüsi meetodid (kahealuselise graafi klasterdamine, intervallgraafi servade ristumiste arvu minimeerimine) on eelkõige sobilikud suvaliste relatsiooniliste süsteemide kinniste hulkade eraldamiseks ja nende korrastamiseks. Seega sobib hästi neid kasutada ka *wordnet*-tüüpi sõnastike hetkeolukorra kajastamiseks.

Positiivne tagasiside Eesti Wordnet'i tegijatelt ja valdajatelt lubab olla kindel, et esitatud meetodid oleksid kasulikud ka teiste keelte *wordnet*-süsteemidele.

Viidatud kirjandus

Flannery, P. Brian; Press, H. William; Teukolsky, A. Saul; Vetterling, T. William 2009. Numerical Recipes in C. The Art of Scientific Computing. South Asia: Cambridge University Press India.

Garey, Michael R.; Johnson, David S. 1979. Computers and Intractability: A Guide to the Theory of NP-Completeness. New York: W. H. Freeman.

Knuth, Donald Ervin 1968. Fundamental Algorithms. The Art of Computer Programming, Vol. 1. Reading, MA: Addison-Wesley.

Kozareva, Zornitsa; Hovy, Eduard 2010. A semi-supervised method to learn and construct taxonomies using the web. – Conference on Empirical Methods in Natural Language Processing. MIT, Massachusetts, USA, 9-11 October 2010, 1110–1118.

Lohk, Ahti; Võhandu, Leo 2011. Eesti sünovara visualiseerimisest. – Ettekanne 10. rakenduslingvistika kevadkonverentsil, 28.-29. aprill 2011, Tallinn.

Niermann, Stephan 2005. Optimizing the Ordering of Tables With Evolutionary Computation. – The American Statistician, 59 (1), 41–46.

Orav, Heili; Kerner, Kadri; Parm, Sirli 2011. Eesti Wordnet'i hetkeseisust. – Keel ja Kirjandus, 1, 96–106.

Vider, Kadri 2001. Eesti keele tesaurus – teooria ja tegelikkus. – M. Langemets (Toim.). Leksikograafiaseminar "Sõna tänapäeva maailmas" / Leksikografinen seminaari "Sanat nykymaailmassa". Ettekannete kogumik. Eesti Keele Instituudi toimetised, 9. Tallinn: Eesti Keele Sihtasutus, 134–156.

Võrgumaterjalid

EuroWordNet http://www.illc.uva.nl/EuroWordNet/ (1.10.2011).

Filosoft OÜ tesaurus http://www.filosoft.ee/thes_et/ (1.10.2011).

Princetoni Wordnet http://wordnet.princeton.edu/ (01.10.2011).

Tartu Ülikooli eesti keele tesaurus (TEKSaurus) ehk Eesti Wordnet http://test.cl.ut.ee/ressursid/teksaurus/ (01.10.2011).

Ahti Lohk (Tallinna Tehnikaülikool), uurimisvaldkond on teksti ja keeruliste suhtesüsteemide uurimine.
ahti.lohk@ttu.ee

Leo Võhandu (Tallinna Tehnikaülikool), uurimisvaldkond on suurte andmesüsteemide süvastruktuuride uurimine.
leov@staff.ttu.ee

STRUCTURAL ANALYSIS OF THE CURRENT ESTONIAN WORDNET

Ahti Lohk, Leo Võhandu
Tallinn University of Technology

Control of any expanding and developing system requires a feedback control mechanism to evaluate the normal trends of the system and also the unsystematic steps. Experience has shown that often more progress is achieved in a field if non-specialists intervene. For this reason the authors (from the field of informatics) suggest special data processing and visualization methods for the Estonian Wordnet. The images presented are suitable as a feedback mechanism for every Wordnet-style linked system, for semantic analysis and examination of regularities. They allow the lexicographer to use an innovative approach to study the hidden structures of the system and make corrections if needed. Several complex data analysis methods have been used (bipartite graph clustering, minimization of interval graph crossing numbers) and the results presented in this article. Positive feedback from the makers and maintainers of the Estonian Wordnet allows us to be sure that the methods presented would be useful for other language Wordnets also.

Keywords: Thesaurus, semantic relations, visualization of Wordnet, ordering of relations, Estonian