TARTU ÜLIKOOL

Arvutiteaduse instituut

Informaatika õppekava

Elina Pankrašin

Eesti Wordneti visualiseerimine

Bakalaureusetöö (9 EAP)

Juhendajad: Sven Aller

Heili Orav

Tartu 2017

Eesti Wordneti visualiseerimine

Lühikokkuvõte: Selle lõputöö eesmärk oli Eesti Wordneti sõnastiku visualiseerimine, kuna visualiseerimine võimaldab kasutajal paremini hoomata keerulist infot, kui see on esitatud graafide või graafikutena. Töö teoreetilises osas antakse ülevaade *wordnet*-tüüpi sõnastiku ajaloost ja ülesehitusest ning visualiseerimisvõimalustest. Praktilises osas tutvustatakse bakalaureusetöö käigus valminud visualiseerimisprogrammi – selle üldkirjeldust, algoritmi, tehnilist poolt ja edasiarendamise võimalusi.

Võtmesõnad:

keeleressursid, visualiseerimine, eesti keel, *wordnet*

CERCS: P175, informaatika, süsteemiteooria

Visualisation of Estonian Wordnet

Abstract:

The purpose of this Bachelor's thesis was to visualise an Estonian Wordnet dictionary, because visualising helps the user to understand difficult information better when it is presented as a graph. In the theoretical part of the thesis, there is an overview of the history and structure and visualisation options of WordNet-type dictionary. In the practical part, there is an overview of the program created for this thesis – its general overview, algorithm, technical solution and possibilities of further development.

Keywords:

language resources, visualisation, Estonian, WordNet

CERCS: P175, informatics, system theory

Sisukord

Sissejuhatus

Sõnastikuks nimetatakse andmebaasi, kuhu kuuluvad sõnad või muud keelelised väljendid ning nende tähendused. Enamasti on sõnastikud tähestikupõhised, kuid mõisteid neis võib jagada ka mingisuguse muu reegli järgi. Tänapäeval keskendutakse üha rohkem igapäevaelu elektroonilisemaks muutmisele. Ka väga paljud sõnastikud on veebi üle kantud.

Veebisõnastike hulka kuulub w*ordnet*, mis on keele leksikaal-semantiline andmebaas. Erinevalt tavaliselt tähestikul põhinevast sõnastikust on *wordnet*'is mõistepõhine ehk mõistet näitab sünonüümihulk. Sõnad on jagatud vastavalt nende tähendusele hulkadesse, mis tähistavad sama mõistet. *Wordnet* on üks peamisi keeletehnoloogia leksikaalseid allikaid.

Eestis hakati oma tesaurust arendama 1995. aastal. Alguses sisaldas Eesti Wordnet (EstWN) inglise keelest tõlgitud baasmõisteid, kuid praeguseks sisaldab see juba 83 500 mõistet.

Antud bakalaureusetöö eesmärk on visualiseerida *wordnet*-tüüpi sõnastik, mis oleks loogilisem ja inimestele arusaadaam kui praegune Eesti Wordneti päringusüsteem TEKsaurus.

Visualiseerimine teeb kergemaks vigade leidmise. Samuti aitab see inimestel paremini mõista sõnade omavahelisi seoseid.

Töö on jagatud nelja peatükki. Esimeses peatükis tutvustatakse keeleressursse. Teises peatükis kirjeldatakse *wordnet*-tüüpi sõnastikku, antakse ülevaade selle ajaloost ning viiakse lugeja kurssi ühega Eesti mõistelistest sõnastikust – Eesti Wordnetist. Samuti kirjeldatakse selles peatükis *wordnet*'i semantilisi suhteid ning tutvustatakse mõned neist. Kolmandas peatükis antakse lugejale ülevaade sõnastike visualiseerimisest ning kirjeldatakse erinevaid sõnastike visualiseerimise võimalusi. Neljandas peatükis keskendutakse valminud programmile – kasutaja saab tutvuda üldkirjelduse, tehnilise teostuse, erinevate katsetuste ja edasiarendamise võimalustega.

Mõistete loetelu

Tabel 1. Mõistete seletused.

Mõiste	Tähendus
leksikosemantiline	sõnatähenduslik (EKI, 2017)
sünohulk	sünonüümirida, mis koosneb üht mõistet väljendavatest samatähenduslikest sõnadest ja sõnaühenditest (Tartu Ülikooli arvutilingvistika uurimisrühm, 2016)
lausung	lause(te)st koosnev kõneakt (EKI, 2017)
süntagmaatilised suhted	seovad lausungi liikmeid (Pajusalu, 2009)
paradigmaatilised suhted	tähistavad lausungi liikmete vahelisi seoseid (Pajusalu, 2009)
tesaurus	sünonüümisõnastik (Prabhat, 2016)
leksikaalne	sõnavaraline (EKI, 2017)
semantiline	tähenduslik (EKI, 2017)
keeleressurss	arvutites loomuliku keele uurimiseks või keeletehnoloogia arendamiseks kasutatav andmekogum, mis on masinloetaval kujul (Tartu Ülikool jt, 2017)
leksikaalne ressurss	sõnastikud, terminoloogilised ressursid, mõistete andmebaasid, sagedusloendid jne (Tartu Ülikool jt, 2017)
leksikaalne andmebaas	keele lekseemide korrastatud kirjeldus (Loos jt, 2003)

1. Keeleressursid

Keeleressursiks nimetatakse keelelist elektroonilist andmekogumit, mis on masinloetav. Selle kasutusaladeks on näiteks keeletehnoloogia arendamine ning arvutites loomuliku keele uurimine (EKRK, 2017).

Eesti Keeleressursside Keskus[1] tegeleb eesti keele digitaalsete ressursside ja tehnoloogiate kogumise, hoiustamise ja kättesaadavaks tegemisega nii teadlastele kui ka muudele huvilistele. Kasutamise mugavuse huvides on juurdepääs digitaalarhiividele ühendatud ning keeletehnoloogia vahendid esitatud arhiveeritud andmeteid kasutava veebiteenusena (EKRK, 2017).

Keeleressursside alla kuuluvad tekstikorpused, kõneandmebaasid, leksikaalsed ressursid, teksti- ja kõnetöötlusvahendid. Lisaks arvestatakse keeleressursiks ka muude keeleressursside administreerimiseks ja töötlemiseks ettenähtud tarkvara (EKRK, 2017). Antud töös kasutatakse leksikaalsete ressursside alla kuuluvat Eesti Wordneti.

Keeleressursse kasutatakse kõne- ja keeletehnoloogias, integreeritud rakendustes ning keeletehnoloogiliste ressursside alal. Neist on abi nii kõnesünteesis ja -tuvastuses, rakendussüsteemide loomisel, erisuguste kõne- ja keeletehnoloogiliste vahendite integreerimisel, tekstitöötluse abivahendites, masintõlkes (EKT, 2017).

[1] http://www.keeleressursid.ee

2. Wordnet

Wordnet'iks nimetatakse keele leksikosemantilist andmebaasi (Miller, Beckwith, Fellbaum, Gross, Miller, 1990). Sõnad on taolises tesauruses jagatud sünonüümide gruppidesse, mida nimetatakse sünohulkadeks. Sünohulka kuuluvad kõik sõnad või sõnaühendid, mis tähistavad ühte ja sama mõistet. Ühes sünohulgas olevad liikmed on kõik samast sõnaliigist. Kui sünonüümigruppi kuulub ainult üks sõna, saab antud mõistet väljendada vaid ühe sõnaga (Orav jt, 2017).

Inglise keele WordNet sai alguse Princetoni ülikoolis aastal 1985. Psühholingvistid eesotsas psühholoogia professori Georg Milleriga hakkasid looma leksikaalset andmebaasi, mille aluseks olid tolleaegsed inimese semantilise mälu teooriad. Neid teooriaid arendati 1960ndate teises pooles, kus leksikon pidi põhinema sõnadel ja nende omavahelistel seostel meie peas ning toetuma mõistetele, mitte tähestikule (Miller jt, 1990).

Algselt loodi *wordnet* psühholoogide ja keeleteadlaste jaoks, ent tänapäeval on ta eelkõige keeletehnoloogiline ressurss (Parm, Orav, 2014). Samuti on see abiks infootsisüsteemides, keeleõppeprogrammides, keeleteaduses ja tehisintellekti rakendustes.

Mõisteline sõnaraamat muutus populaarseks eelkõige seetõttu, et tekkis vajadus muuta loomuliku keele mõisteseosed arusaadavaks ka arvutisüsteemidele. Arvuti peab keeleandmetele põhinedes teksti kohta erinevaid järeldusi teha ning *wordnet* arvutiressursina on oluliseks abiks (Parm, Orav, 2014).

Elektrooniline sõnaraamat kaotab ruumilised piirangud – see on palju odavam ja vähemahulisem kui paberkujul variant. Lisaks on kasutajal elektroonilisel kujul palju mugavam infot nii hoomata kui ka otsida (Langemets, Kallas, 2014).

Wordnet-tüüpi leksikosemantilisi andmebaase on rohkem kui 60 keeles ning neid luuakse üha juurde. Mõnedel keeltel on selliseid andmebaase ka mitu (EKT, 2014).

2.1 Eesti Wordnet

Eesti üldkeele masinloetava tesauruse loomisega tehti algust Tartu Ülikooli arvutilingvistika uurimisrühmas juba aastal 1995. Eestis oli palju nii arvutiga koostatud kui ka arvutisse viidud leksikone, kuid polnud sellist arvutisõnastikku, mis oleks erinev tavalisest sõnaraamatust nii struktuuri kui ka eesmärgi poolest (Orav, Kerner, Parm, 2011).

Üheks olulisemaks erinevuseks tavalise sõnaraamatu ja *wordnet*'i vahel on see, et sõnad pole jaotatut tähestikupõhiselt, vaid hoopis sõnaliikide ja tähenduste järgi (Miller jt, 1990). Nagu *wordnet*'ides üldiselt, leiab ka EstWN-ist lisaks sõnade tähendusetele ka tähendustevahelised seosed. Seetõttu kasutatakse seda andmebaasi ka keeleteaduslikes uurimistöödes ning raallingvistilistes rakendustes (Orav, Kerner, Parm, 2011).

1996. aastal liitus uurimisrühm projektiga EuroWordNet, mis seadis endale eesmärgiks koostada mitmekeelne *wordnet*-tüüpi tesaurus. Eesti Wordneti kallal käib töö siiani. Kõik mõisted on ühendatud keeltevahelise indeksi abil Princetoni WordNeti ja ka teiste keelte tesauruste mõistetega – see aitab mõisteid erinevates keeltes omavahel võrrelda.

Esimesteks mõisteteks oli baasmõisted, mis tõlgiti inglise keelest eesti keelde. Eesti Wordnet sisaldab nimi-, tegu-, omadus- ja määrsõnu ning mitmesõnalisi ühendeid. Eesti keele põhisõnavara on Eesti Wordnetis nüüdseks juba olemas (Tartu Ülikooli arvutilingvistika uurimisrühm, 2016). 2017. aasta aprilli seisuga sisaldab EstWN rohkem kui 83500 mõistet (EKT, 2017).

Eesti Wordnet on üks olulisemaid eesti keeleressursse. See on lisaks Andrus Saareste „Eesti keele mõistelisele sõnaraamatule" ainuke mõistepõhine arvutileksikon (Parm, Orav, 2014).

2.2 Semantilised seosed

Semantilised suhted tähistavad sõnade, fraaside ja lausete vahel olevaid seoseid. Sõnadevahelisi suhteid tähistavad sünonüümia, antonüümia, homonüümia, polüseemia, metonüümia (Zapata Becerra, 2000).

Semantilised seosed võivad olla kas paradigmaatilised või süntagmaatilised (Pajusalu, 2009):

- Paradigmalisteks nimetatakse sisuseoseid. Need tähistavad objektide ja nähtuste vahel esinevaid suhteid tegelikkuses. Sinna alla kuuluvad näiteks süno-, hüper-, hüpo- ja antonüümia.
- Süntagmaatilised suhted peegeldavad seda, kuidas on lausungi elemendid üksteisega seotud. Näiteks kollokatsioonid ja idioomid on süntagmaatilised suhted.

Sõnavõrgustik põhineb sõnade paigutusel meie ajus. WordNeti loomise käigus viidi läbi mitmeid psühholoogilisi katseid, mis näitasid seda, et kõnelejad organiseerivad oma mõistelisi teadmisi hierarhilisel moel. Samuti saadi teada, et mõisteliste teadmiste taastamiseks kuluv aeg on otseselt seotud sellega, kui palju hierarhiaid on kõnelejal vaja läbida teadmiseni jõudmiseks (Wiley, 1972).

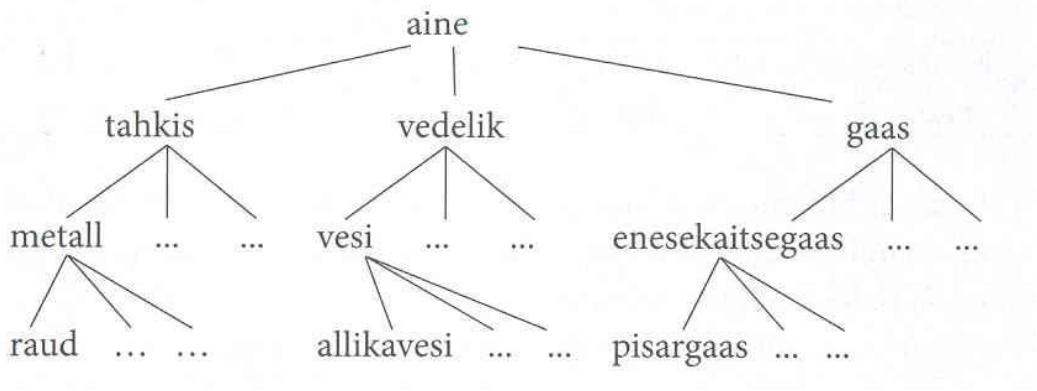
Seni on *wordnet*'i loomisel tuginetud inglise keele vastavale andmebaasi ülesehitusele (Orav jt, 2011). Omakeelset lähenemist see aga ei toeta, seetõttu proovitakse lisaks tesauruse mahukuse suurendamisele leida just eesti keelele rakendatavaid semantilisi seoseid. Välja võib tuua selle, et määrsõnade ja muude sõnaliikide vahel on selliseid tuletusseoseid, mida pole võimalik praeguse tesauruse formaadiga võimalik kujutada – näiteks kui lisada määrsõnast tuletatud omadussõnale määrsõna sõnalõpp –lt (määrsõna *praegu* → omadussõna *praegune* → määrsõna *praeguselt*). Hetkel kasutusel olev tesauruse vorming lubab antud juhul ühe sõnahulga alla määrata lekseeme nii, et määrsõnad *praegu* ja *praeguselt* määratakse samasse sünohulka ning ühendatakse tuletussuhte abil omadussõnaga *kohene*. Seda võiks aga hoopiski tähistada nn „tagasituletuse leksikaal-semantilise suhtega".

Mõistetevaheliste semantiliste suhete kindlaks tegemisel on olulised kaks erisugust näitajat. Arvestada tuleb nii eesti keele sõnavara eripäraste suhetega teoreetilises mõttes kui ka tesauruse olulisusega keeletehnoloogiliste rakenduste jaoks. Viimane on oluline, sest mida detailsem ja mahukam on suhtevõrgustik, seda parem on arvutiressurss (Orav jt, 2011).

Järgnevalt kirjeldataksegi semantilisi suhteid, mis sõnavõrgustikus olulist rolli mängivad.

2.2.1 Hüpero- ja hüponüümia

Hüponüümia tähistab tähendustevahelisi soo-liigi seoseid. Hüponüüm tähistab alammõistet ning hüperonüümiks nimetatakse sõna ülemmõistet (Heinmets, 2017).



Joonis 1 Alam- ja ülemmõisted. (Heinmets, 2017).

Joonisel 1 on kujutatud mõiste „aine" hierarhiat. Siinkohal on „aine" nii „tahkise", „vedeliku" kui ka „gaasi" hüperonüüm. „Metall" on nii hüponüüm (sõna „tahkis" suhtes) kui ka hüperonüüm („raua" suhtes).

Hierarhias kujutatakse kõrgemal laiema tähendusega mõisteid, allpool aga kitsama tähendusega mõisteid.

2.2.2 Rollisuhted

Osalus- ja rollisuhted tähistavad enamasti tegu- ja nimisõnade vahel olevat seost, mis tähistab seotust tegusõnaga. Mõnikord võivad need seosed esineda ka teiste sõnaliikide vahel (Payne, 2007).

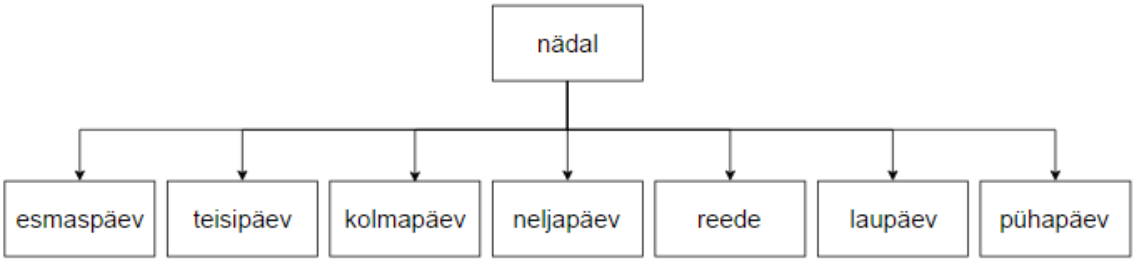
Tabel 2. Erinevaid rolli- ja osalussuhteid.

Osalussuhe	Tähendus	Näide
Agent	Tegevuse teostaja	**Ettekandja** tõi supi.
Kogeja	Sündmuse poolt mõjutatu	**Sille** kuulas muusikat.
Jõud	Sündmuse mittetahtlik põhjusaja	**Tsunaami** hävitas linna.
Patsient	Sündmuses osaleja; võib olla kellestki/millestki mõjutatud	Tuul lõhkus **linnupesa**.
Põhjus	Sündmuse põhjustaja	**Kuna Jukul oli õppimata**, ei suutnud ta kontrolltööd edukalt sooritada.
(Asu)koht	Sündmuse toimumise paik	**Aias** kasvavad puud
Vahend	Sündmuse läbiviimise vahend	Mari lõikas **noaga** kooki.

Tabelis 1 on ülevaate andmiseks välja toodud mõned osalus- ja rollisuhted.

2.2.3 Osa-terviku suhted

Meronüümia tähistab mõistetevahelisi osa-terviku suhteid. Tervikut tähistavat sõna nimetatakse holonüümiks, osa väljendavat sõna aga meronüümiks (Erelt, Erelt, Ross, 2017).



Joonis 2 Osa-terviku suhte (Erelt, Erelt, Ross, 2017).

Nagu on näha joonisel 2, on nädalapäevad meronüümid ja nädal ise holonüüm. Käsi tähistab tervikut ja sõrmed on osa sellest.

3. Sõnastike visualiseerimise vajalikkus

Selleks, et sõnastikud veelgi mugavamad oleks, on hakatud neid ka visualiseerima. Esiteks võimaldab see kasutajal infot paremini hoomata. Inimese aju aktsepteerib suurt kogust infot paremini, kui see on esitatud graafide või graafikutena, mitte tabelite ja raportitena. Teiseks muudab see keerulisema info arusaadavamaks ning kasutatavamaks (Few, 2015).

Viimasel ajal on populaarseks muutunud sõnastike visualiseerimine. Lisaks mugavdatakse leksikone erinevate graafikute ja sõnapilvede abil, kuna need võimaldavad paremini hoomata sõnadevahelisi seoseid. Samuti lisatakse neisse pilte ja fotosid (Langemets, Kallas, 2014).

3.1 WordTies

WordTies[2] on veebipõhine kasutajaliides, mida arendati Kopenhaageni Ülikoolis ükskeelsete *wordnet*'ide loomiseks ning nende omavaheliseks kõrvutamiseks META-NORD projekti käigus[3]. Tegemist on pilootprojektiga (Pedersen jt, 2013).

WordTiesis kujutatakse eri suhteid eri värvi joontega, mis võimaldab saada hea ülevaate *wordnet*'i üldstruktuurist. Semantiliste suhete jälgimiseks loodud kasutajaliides on üles ehitatud ükskeelsele brauserile AndreOrd, mis laseb erinevalt teistest veebilehitsejatest vaadelda semantilisi suhteid graafiliselt. Vajutades graafil seotud sünohulgale saab *wordnet*'is dünaamiliselt ringi liikuda (Pedersen jt, 2013).

Hetkel on WordTiesi eestikeelne variant seotud teiste keeltega vaid 1000 mõiste kaudu. Need tuhat mõistet on poolautomaatselt kompileeritud kõige enam kasutatavate sõnatähenduste loendist mida nimetatakse „core *wordnet*'iks" (Fellbaum ja Tengi, 2017). Samuti on eestikeelne andmebaas vananenud – see on endiselt 2013. aasta seisuga (otsimisvõimalus on ligi 57 tuhandele eestikeelsele mõistele) – ja kuna projekt on lõppenud, pole andmeid enam võimalik uuendada.

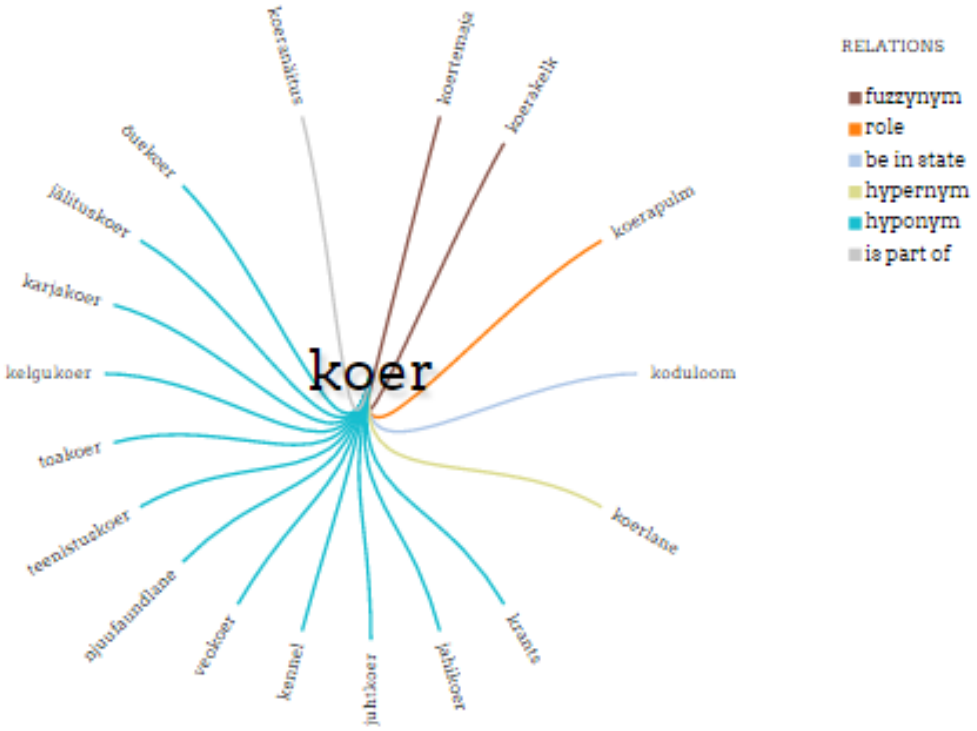
Illustreerimiseks on allpool joonis 3, millel on kujutatud sõna „koer" otsing eestikeelses WordTiesis ning selle joondamine inglise, taani-, soome-, poola- ja rootsikeelsete vastetega. Sel moel kergendab veebi kasutajaliides *wordnet*'ide võrdlemist ja hindamist.

[2] http://wordties.cst.dk (3.05.2017)
[3] http://www.meta-nord.eu (3.05.2017)

koer

NOUN

(koer) peamiselt hundist põlvnev koduloom



View **relations** or concept hierarchy

Synonyms

peni

Alignment

The English Wordnet has concepts that correspond to the Estonian:

→ *(English)* dog, domestic dog, and Canis familiaris: *a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds*

→ *(Danish)* hund, koter, vovse, and vovhund : *pattedyr som har god lugtesans og hørelse, og som ...*

→ *(Finnish)* Canis familiaris and koira : *"a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; ""the dog barked all night"""*

→ *(Polish)* pies :

→ *(Swedish)* hund :

Joonis 3 Eestikeelse sünohulga võrgustik mitmekeelsete võrdustega inglise, taani, soome, poola ja rootsi keeles

Vajutades lingile jaotise *Alignment* juures liigub kasutaja edasi valitud keele *wordnet*'i ning saab uurid sama mõistet teises keeles. Uurida on võimalik nii suhteid kui ka mõistete hierarhiat (*concept hierarchy*).

Nagu on näha jooniselt 3, on hetkel enamus kasutajaliidesest ingliskeelne. Vaid mõisted ja nende definitsioonid on kasutatavas keeles.

Joonis 4 Soome ja eesti wordnetide taksonoomilise struktuuri võrdlus sõnaga "liikuma"

Joonis 4 võrdleb eesti ja soome *wordnet*'ide vahelist taksonoomilist struktuuri 'liikuma' kohta.

3.2 Poola Wordneti visualiseerimine

Oma visualiseerimisprogrammi on teinud ka poolakad ning siinkohal on sellest lühitutvustus. Poola Wordnet'i (plWN[4]) visualiseeriti WordNet Editori[5] abil, mis on WordNet Solutioni süsteemi osa. WordNet Solutioni arendusega tegelevad Gdański Tehnikaülikooli arvuti arhitektuuri, elektroonika, telekommunikatsiooni ja informaatika teaduskonnad. Antud süsteem on spetsiaalselt mõeldud *wordnet*'i välissüsteemides muutmiseks, visualiseerimiseks ja integreerimiseks (Szymański, Chodor, 2007).

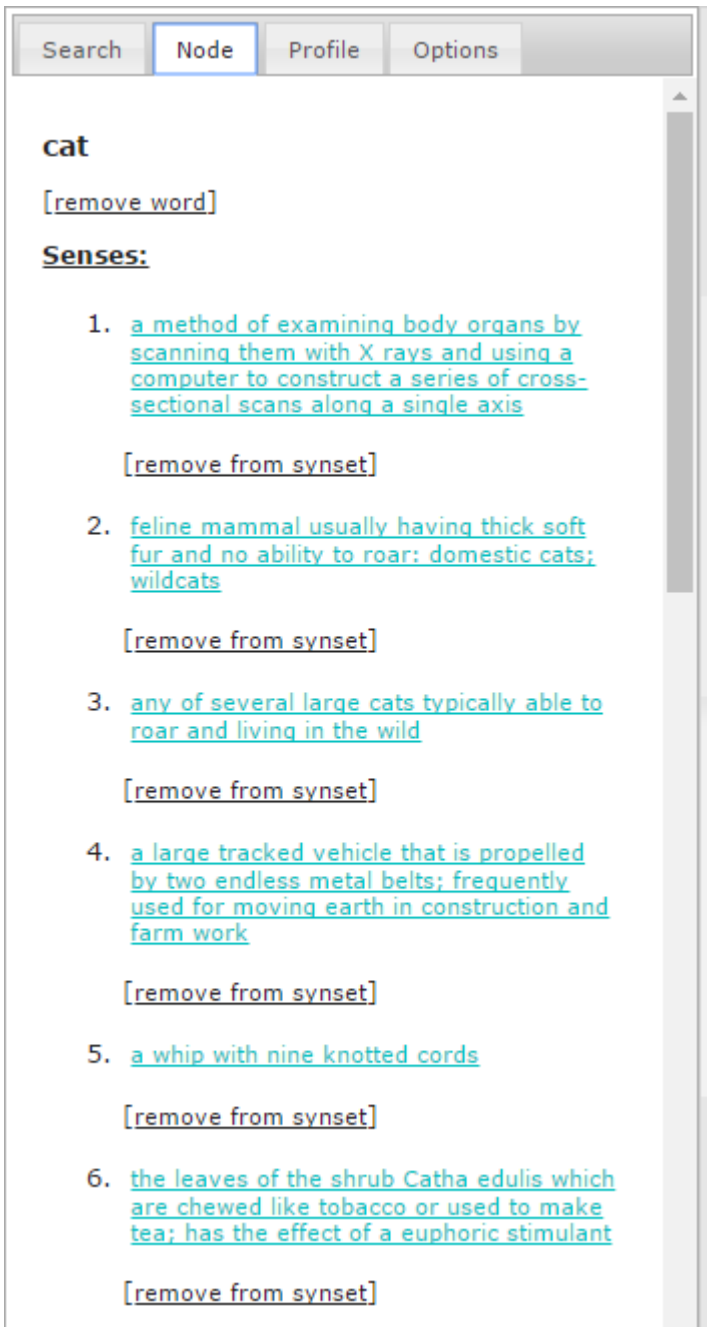
[4] http://wordventure.eti.pg.gda.pl/wne/wne.html (11.05.2017)
[5] http://wordventure.eti.pg.gda.pl/

Kuvatõmmis kasutajaliidesest on toodud joonisel 5. Otsingut teostati ingliskeelse sõnaga „cat“. Kõigepealt pakutakse kasutajale valida, millist sõna või fraasi ta täpsemalt näha soovib. Valikus on nii sõna „cat“ eraldi kui ka fraaside osana.

Joonis 5 Otsing sõnale "cat" Poola Wordneti visualiseerimise liideses[6]

[6] http://wordventure.eti.pg.gda.pl/wne/wne.html (11.05.2017)

Lisaks otsinguaknale on liideses ka aken, kust leiab otsitud sõna erinevad tähendused (joonis 6). Oletame, et soovitakse lähemalt uurida kassi kui looma. Valides teise tähenduse, hoiatab meid süsteem kõigepealt, et ekraanile ilmub lisa 64 sõnapilve. Nõustudes sellega, uuendatakse antud akna sisu.

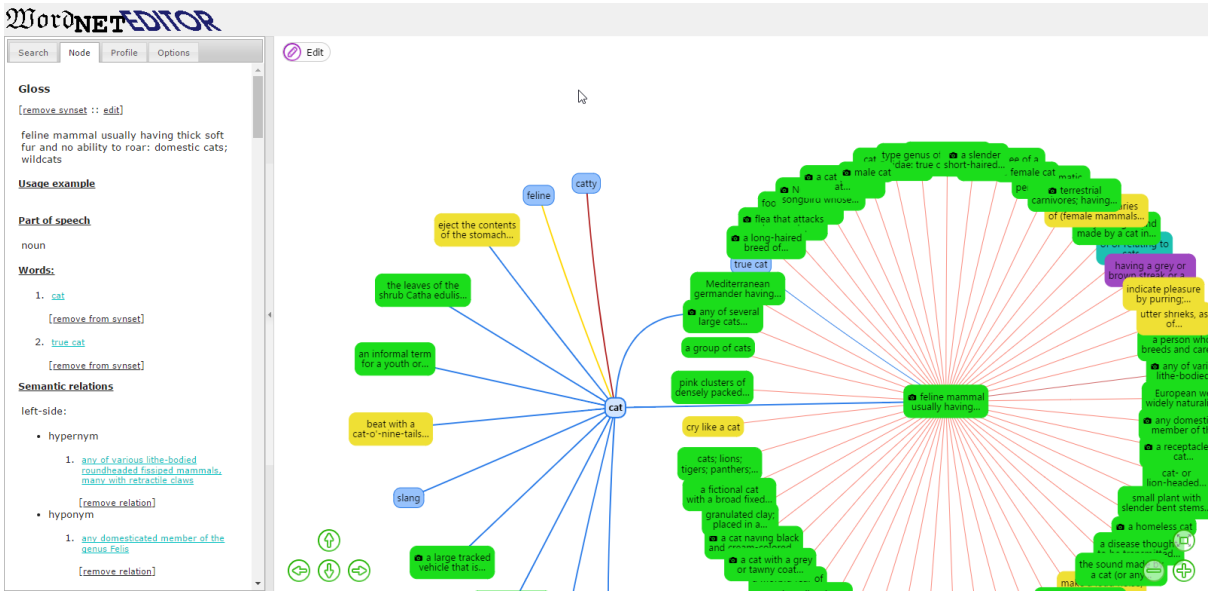


Joonis 6 Sõnade tähenduste aken[7]

[7] http://wordventure.eti.pg.gda.pl/wne/wne.html (11.05.2017)

Joonisel 7 on näha, et ekraanile ilmusid otsitava sõna erinevad tähendused ning vasakul pool saab vaadata otsitud sõna tähendust, sõnaliiki, semantilisi suhteid. Võimalusel esitatakse ka näide, kuidas sõna lausetes kasutatakse.

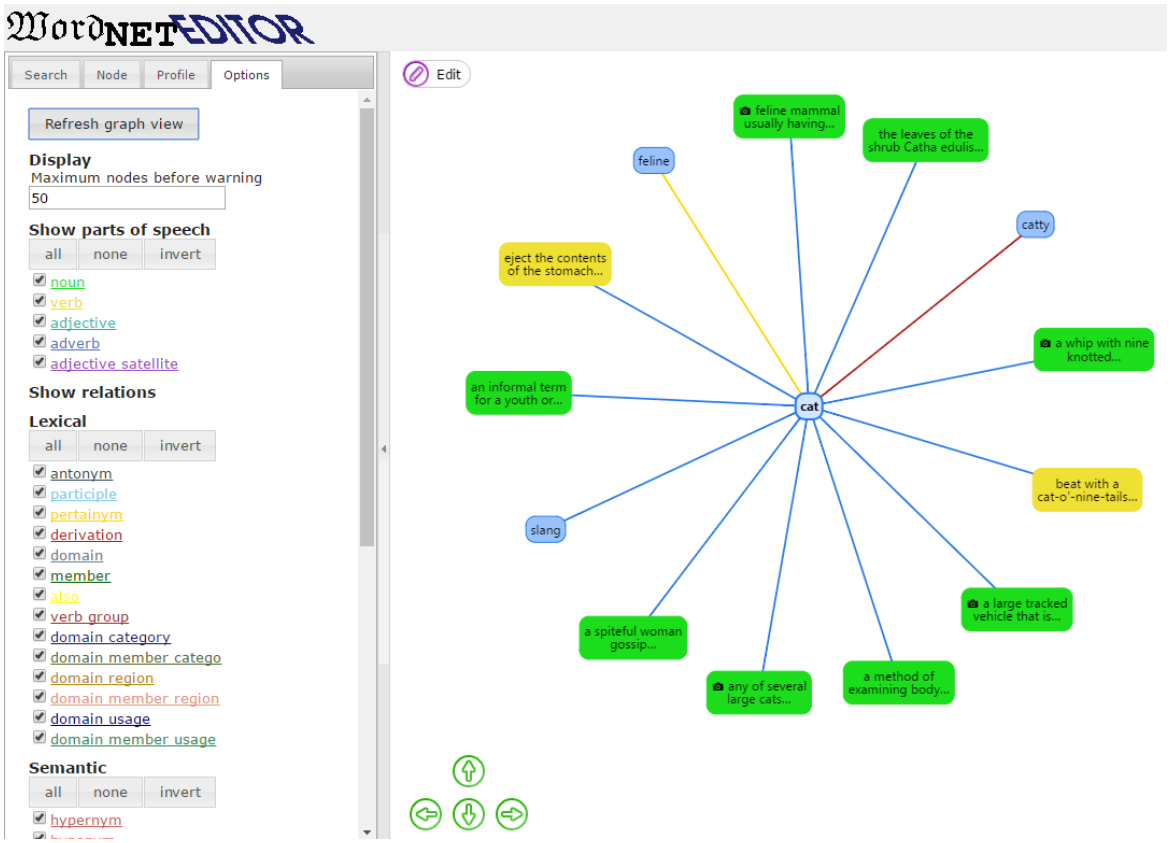
Rõhutamiseks on vasakul pool olevad lingid teist värvi. Läbi nende on võimalik tekitada sõnapilvi juurde ja liikuda *wordnet*'is aina sügavamale.



Joonis 7 "Kass" tähenduses "koduloom" plWN-is.[8]

Hetkel on valitud aken „Options" („Seaded"), kust saab valida, milliseid suhteid kasutaja soovib hetkel uurida. Seoseid on võimalik valida nii sõnaliigi järgi, leksikaalseid kui ka semantilisi. Tuleb tähele panna, et pärast iga muudatuse sisseviimist peab vajutama nuppu „Refresh graph view", mis uuendab graafi.

[8] http://wordventure.eti.pg.gda.pl/wne/wne.html



Joonis 8 Poola Wordneti visualiseerimisprogramm Wordnet Editor[9]

Poola Wordneti visualiseerimisprogrammil on nii häid kui ka halbu külgi.

Positiivsetest võib välja näiteks selle, et kogu info on eraldi kastis vasakul pool ära toodud. See teeb kasutajale programmi kasutamise kergemaks, sest on võimalus liikuda erinevate akende vahel sobiva info saamiseks.

WordTiesiga võrreldes on Wordnet Editoril hea see, et kasutajal on võimalik viimases aknas ise valida, milliseid suhteid ta näha soovib. Lisaks saab valida sealt korraga kas kõik, mitte ühtegi või hoopiski vastupidised suhted neile, mida kasutaja eelnevalt otsis. Siiski peab välja tooma, et soovitud muutuste rakendamiseks peab ülevalt valima „Refresh graph view“

Osade mõistete juurest võib leida ka pilte. Ühest küljest aitab see kasutajal keerulisemaid ja tundmatuid mõisteid paremini mõista. Teisest küljest on osadel sõnaseletustel pildid, mis enam ei ilmu, kuna need on kas oma originaalsetest asukohtadest eemaldatud või mujale liigutatud.

Üheks suureks miinuseks võib mainida seda, et otsides uut mõistet, jäetakse eelmine mõiste koos kõigi suhetega, mida kasutaja eelnevalt avas, alles. See suurendab oluliselt mälukasutust.

[9] http://wordventure.eti.pg.gda.pl/wne/wne.html

Samuti on kasutajal väga raske hoomata kogu infot, mis ekraanile ilmub. Osa mõisteid kaob ära, kui kujutist suurendada; välja suumides ei ole jällegi võimalik kõike välja lugeda. Siiski annab programm kõigepealt hoiatuse, enne kui hakkab kõiki valitud suhteid laadima.

4. Programmi ülevaade

Lõputöös valminud programmi[10] eesmärk oli esitada Eesti Wordneti andmed visualiseeritud kujul, mis oleks abivahend nii *wordnet*'i koostajaile kui ka muidu huvilistele. Programm lubab kasutajal vaadelda erinevaid EstWN-i mõisteid ning nende suhteid teiste mõistetega.

Praegusel hetkel on võimalik otsida 77878 mõistet, st 125 646 sõna.

Eestikeelne WordTies, mis on seni ainuke EstWN-i visualiseeritud versioon, on vananenud andmetega. Sellest lähtuvalt tekkis vajadus sõnastiku visualiseerimise järele, mis sisaldaks kõiki mõisteid ning võtaks neid otse EstWN-i viimase versiooni failist. Kui andmebaasi täiendatakse, siis täieneb pärast vastavat eeltöötlust ka visualiseeritud variant. Siinse töö jaoks loodud variant sisaldab rohkem mõisteid, mis teeb ta kasutajale väärtuslikumaks.

Programmis saab valida, mis suhteid täpsemalt esitatakse. Selline otsus sai tehtud, kuna semantilisi suhteid on väga palju ning need ei mahuks korraga ekraanile. Lisaks tekitaks see kasutajas rohkem segadust ega läheks sellest lähtuvalt kokku antud töö eesmärgiga.

Seosed on märgitud erinevate värvidega. Süno- ja antonüümiat ning hüpero- ja hüponüümiat tähistab teistest suhetest eristamiseks punktiirjoon, kuna tegu on *wordnet*'i peamiste suhetega.

Lisaks on WordTiesi enamik kasutajaliideseset ingliskeelne, mis ei pruugi olla igale tavakasutajale arusaadav. Autori loodud programm on eestikeelne. Osad keerulisemad seosed (nt hüpero- ja hüponüümia) muudeti tavainimesele arusaadavamaks (vastavalt alam- ja ülemmõiste).

Otsitava sõna seosed on paigutatud astmeliselt, mis teeb nende vaatlemise kasutajale mugavamaks. WordTiesis on kõik sõnad erineva nurga all, mis takistab kohati nende loetavust (nt kui sõna on 90-kraadise nurga all).

Mõistete vaheline liikumine on samuti mugavam tänu ajaloole, mis salvestab kõik kasutaja läbitud sõnad. Ükskõik millisele sõnale vajutades saab kasutaja naasta eelnevalt vaadatud sõna juurde.

4.1 Üldkirjeldus

Valminud programm pakub võimalust otsida erisuguseid sõnu ja pakub kasutajale otsitavate sõnadega seotud teisi mõisteid (näiteks läbi süno- ja antonüümia, hüpo- ja hüperonüümia).

[10] http://prog.keeleressursid.ee/EstWNvis/

Programmi kasutamine on väga lihtne. Kuna fail „dictionary.txt“, millest infot laetakse, on üsna mahukas, siis kõigepealt tuleb ära oodata, kuni fail alla laetakse. Kuni faili laetakse, on ekraanil pilt, mis kaob, kui programm on kasutamisvalmis. Seejärel tuleb sisestada soovitud sõna otsingusse.

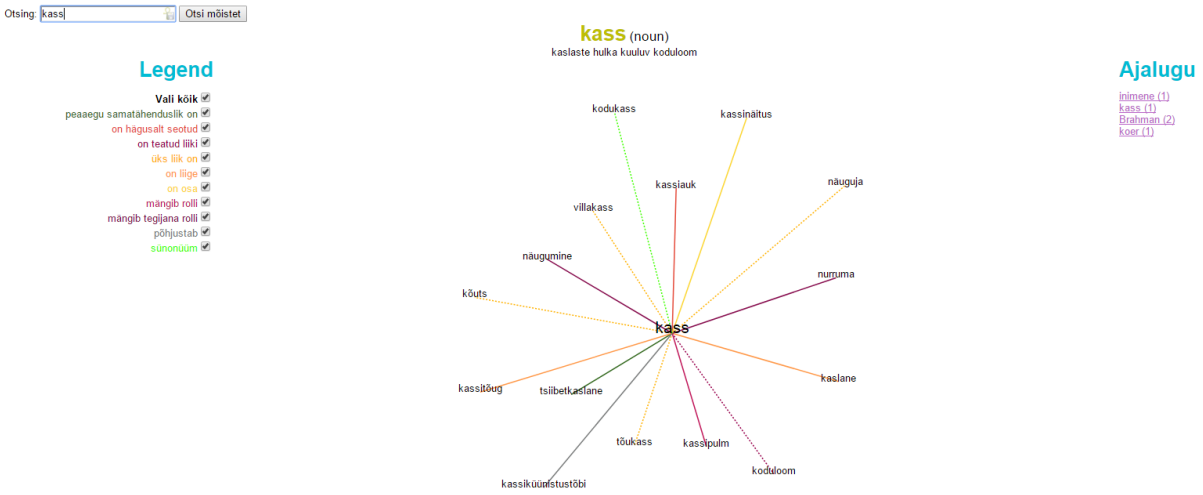
Kui otsitud sõnal on mitu tähendust, suunatakse kasutaja kõigepealt erinevate tähenduste lehele, millest saab valida sobiva. Seejärel laaditakse lehele graaf kõikide sellel sõnal olevate mõistetega. Joonisel 9 on teostatud otsingut sõnaga „koer“. Nagu näha, on antud sõnal 3 erinevat tähendust, millest sobival klikkides suunatakse kasutaja just selle tähenduse seoste graafilisele kujule.

Joonis 9 Sõna „koer“ otsing

Topeltklikkides suvalisel mõistel saab omakorda vaadata seda mõistet koos seostega. Läbikäidud sõnad ilmuvad paremal asuvasse tulpa „Ajalugu“, et vajadusel saaks uuesti neid uurida.

Keskmise ehk antud hetkel otsitava sõna peale klikkides suunatakse kasutaja kõikide tähenduste nimistusse, kui sel sõnal on üle ühe tähenduse. Kui sõna on ainult ühes tähenduses, laaditakse graaf uuesti.

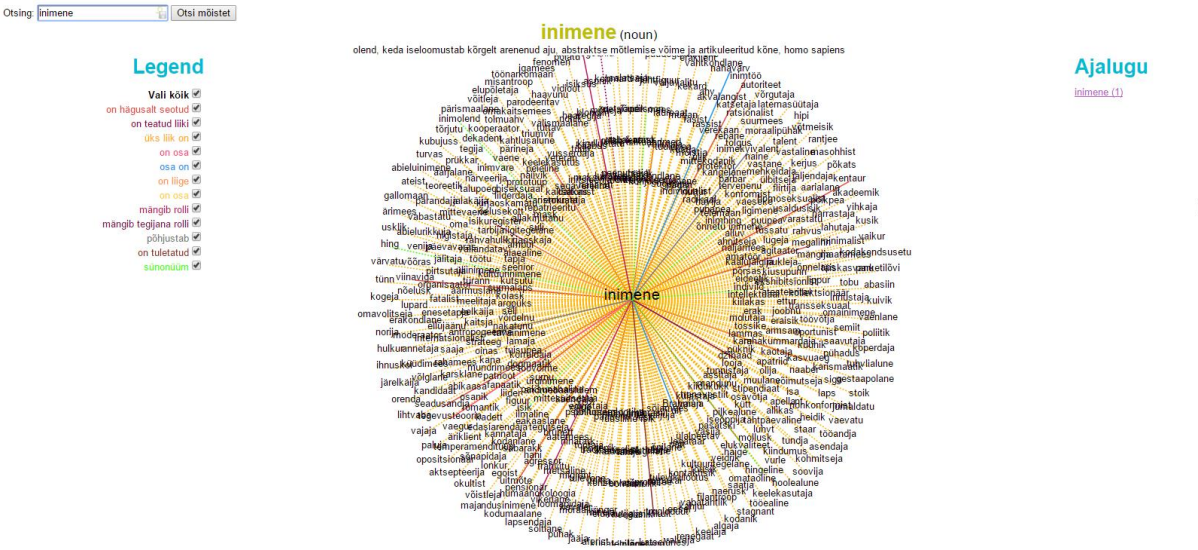
Joonisel 10 teostati otsingut sõnaga „kass“. See sõna on ainult ühes tähenduses ning kuna seoseid on üpriski vähe, on võimalik neid ekraanil normaalselt jälgida.



Joonis 10 Sõna "kass" otsing

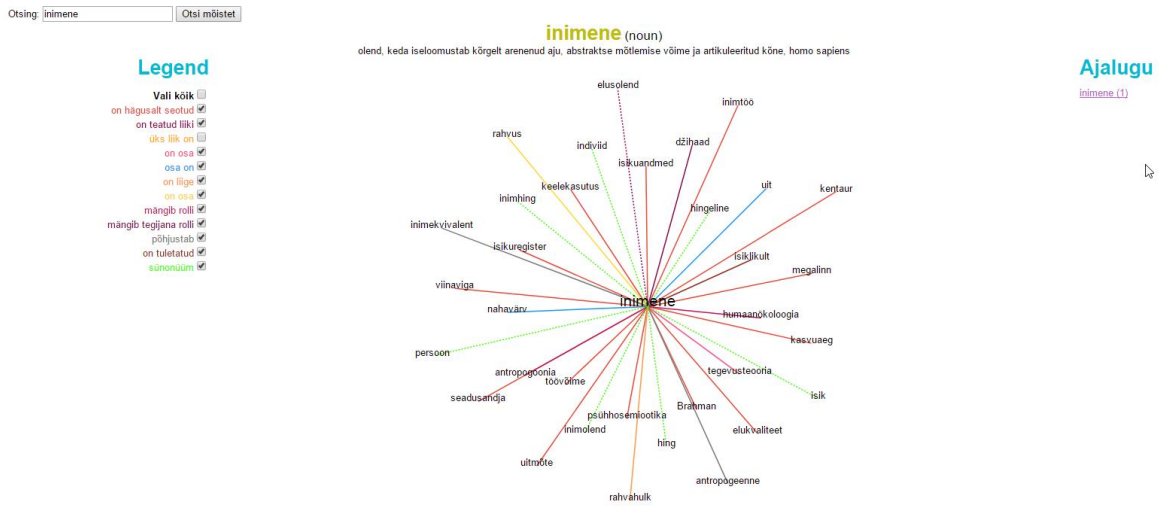
Võib ette tulla olukordi, kus otsitaval mõistel on väga palju erisuguseid seoseid. Joonisel 11 on teostatud otsingut sõnaga „inimene“. See on üks rikkaimast sõnadest Eesti Wordnetis – „inimesel“ on üle 300 alammõiste.

Taoliste sõnade puhul tuleb kasuks vasakul asuv „Legend“, mis lubab kasutajal valida just teda huvitavad mõisted. Kui kasutaja on oma valiku teinud, laaditakse kohe uuenenud graaf.



Joonis 11 Sõna "inimene" otsing koos kõigi alammõistetega

Joonisel 12 välistati graafil „inimese“ alammõisted. Kuna eelmisel graafil on esitatud liiga palju seoseid, võib kasutajal olla raske sellest aru saada. Pärast osade seoste välja jätmist muutus pilt kohe palju arusaadavamaks.



Joonis 12 Sõna "inimene" otsing ilma alammõisteteta

Antud programmis on kasutajaliides erinevalt WordTiesist eestikeelne ja kasutajasõbralikum. Paljude suhetega sõnade puhul on väga kasulik see, et osad suhted saab graafilt ära jätta. Selle abil saab vajaliku suhte üles leida ning n-ö teistest kõrvale viia.

4.2 Tehniline teostus

Valminud programmi tööks kasutati erinevaid algoritme nii eeltöötluseks kui ka programmi enda tööks.

4.2.1 Eeltöötlus

Programmi loomiseks kasutati Eesti Wordneti viimast versiooni (kb73). Esialgne fail on üle 128MB suur. See sisaldab väga palju erinevat infot, mida antud töös ei kasutata ning sellest tingituna sai loodud eraldi „dictionary.txt“ fail, mis sisaldab tööks vajalikku infot. Sõnastikufaili ümbertöötlemiseks kasutati programmeerimiskeelt Python. Antud keel meenutab inglise keelt, seega on selles lihtne koodida ning ta on arusaadav. Samuti on selles väga mugav teha tekstitöötlust (erinevalt näiteks Javast).

Iga uus sõna algab EstWN-i failis 0-ga. Kui programm leiab 0-i rea alguses, siis alustatakse uue sõna lisamist. Esialgu lisatakse uude sõnastikufaili eelnev sõna, kui see on olemas

(esimesel sõnal puudub eelnev sõna). Iga sõna kohta salvestatakse tema sünonüümid, mille kohta omakorda definitsioonid ja tähendused, ning seosed teiste sõnadega.

Sõnad salvestatakse „dictionary.txt“ faili JSON-formaadis.

```
"minema": [{
        "s": 12,
        "p": "verb",
        "l": [
          "3|4|sobima",
          "4|2|mahutama",
          "7|2|mahtuma"
        ],
        "d": "mõõtmetelt sobima"
      },
      {
        "s": 7,
        "p": "verb",
        "l": [
          "3|1|edenema",
          "7|1|kuluma",
          "7|1|mööduma",
          "7|9|kaduma"
        ],
        "d": "aja kohta"
      },
```

Joonis 13 Sõna "minema" failis "dictionary.txt"

Joonisel 13 on kujutatud üht osa failist „dictionary.txt“. Faili mahu vähendamise huvides on võtmed tehtud ühetähelisteks.

„S“ (sense) tähistab sõna tähendust, „p“ (part of speech) sõnaliiki“, „l“ (links) ühendusi ning „d“ (definition) seletust. Ühendused on samuti esitatud kompaktsel kujul – sõne on jaotatud kolmeks, eraldatuna püstkriipsudega. Esimene tähistab ühenduse tüüpi (sünonüüm, antonüüm jne), teine sõna tähendust ja kolmas sõna ise.

4.2.2 Visualiseerimisprogramm

Visualiseerimisprogramm kasutab programmeerimiskeelt JavaScript ning JavaScripti teeki D3.js.

Kliendipoolset veebirakendust pole võimalik teha muu keelega kui JavaScript, kuna väga paljud veebilehitsejad ja operatsioonisüsteemid toetavad just seda keelt. Näiteks Flash ja Java ei ole toetatud kaasaegsete brauserite poolt. Visualiseerimiseks on JavaScriptis olemas teek D3.js, mis võimaldab andmeid veebrauseris visualiseerida dünaamiliselt ja interaktiivselt.

Erinevalt teistest teekidest võimaldab antud teek lõpptulemust väga palju muuta ja manipuleerida (Bostock, 2012).

Kui kasutaja sisestab otsingukasti soovitud sõna, siis kõigepealt laaditakse sisse failist „dictionary.txt“ JSON-andmed JavaScripti objektiks. Kui URL-is on määratud sõna parameetritega *word* ja *sense*, siis laaditakse vastava sõna graaf. Parameetri *sense* puudumisel suunatakse kasutaja sõna tähenduste nimekirja lehele. Juhul, kui puudub parameeter *word*, on võimalik kasutada ainult otsingukasti. Ühegi sõna graafi ei ilmu.

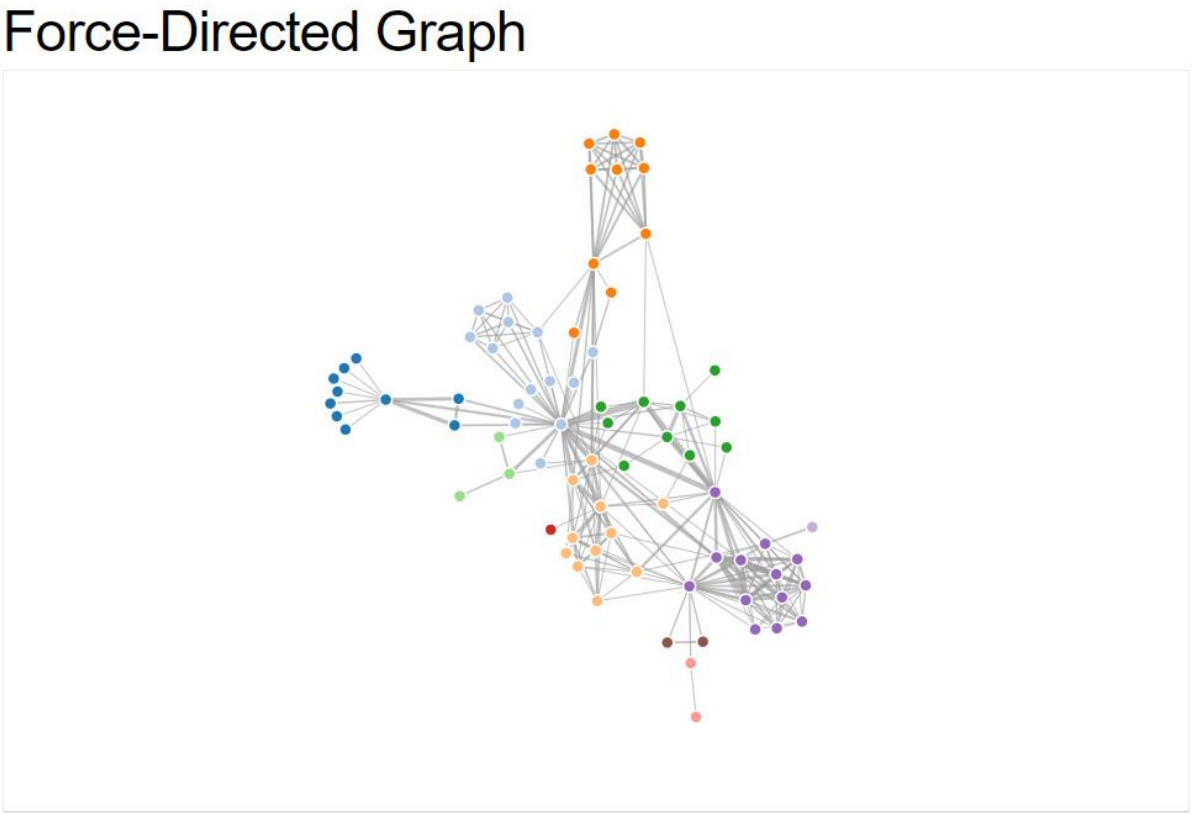
Sõna laadimisel, kui ei ole määratud sõna *sense*'i ja sõnal on ainult üks tähendus, kuvatakse kohe otsitava sõna graaf. Juhul, kui tähendusi on mitu ja *sense* puudub, kuvatakse sõna tähenduste nimekiri ning kasutaja saab valida sobiva tähenduse. Sõna laaditakse andmeobjektist ja töödeldakse graafi joonistamiseks vajalikult kujule.

Graafi joonistamiseks on tarvis luua massiivid graafi tippude (*nodes*) ja tippudevaheliste servadega (*links*). Igast sõnaühendusest, mis on andmeobjektis määratud, luuakse kaks uut objekti: graafi tipp (seotud sõna) ja graafi serv (ühendus laaditud sõna ja temaga seotud sõna vahel). D3.js abil tekitatakse SVG-elemendid, mis vastavad sõnadele ja nendevahelistele seostele. Graafi füüsika simuleerimiseks kasutatakse D3 *forceSimulation* funktsiooni.

Otsitud sõna lisatakse otsinguajaloo massiivi ning ka veebilehitseja ajalukku (`window.history`). Lehelt eemaldatakse eelneva sõna graaf, sõna info sätitakse sobivalt lehele (sõna ise, selle definitsioon, liik). Enne graafi kujutamist filtreeritakse välja seosed, mis pole legendis valitud.

Käivitatakse simulatsioon. Valitud ühenduse tüüpidest ehitatakse legend ning kuvatakse lehele. Kui legendis muudetakse valikut, laetakse sõna täielikult uuesti. Kui kasutaja sisestab sõna, mida pole olemas, kuvatakse ekraanile teade „Sõna ei leitud“.

Visualisatsiooni aluseks võeti Mike Bostock'i loodud Force-Directed Graph demo, mis on kujutatud joonisel 14.



Joonis 14 Mike Bostock'i Force-Directed Graph[11]

Graafi tippe tähistavad värvilised ringid asendati sõnadega. Lisati funktsionaalsus graafi servade stiliseerimiseks vastavalt ühenduse liigile. Graafi keskmesse paigutati vaid üks sõna, millel on oma kindel stiil, ning mis on ühendatud kõigi teiste sõnadega. Samuti lisati seotud sõnade n-ö „astmeline kaugus“ – sõnad jaotatakse aina suurema raadiusega ringidele selleks, et parandada loetavust rohkemate ühenduste puhul. Topeltklikkides suvalise sõna peale laaditakse see sõna keskmise sõnana. Keskmise sõna peale topeltklikkides kuvatakse kõik selle sõna tähendused, kui neid on rohkem kui üks.

4.2.3 Kujundus ja struktuur

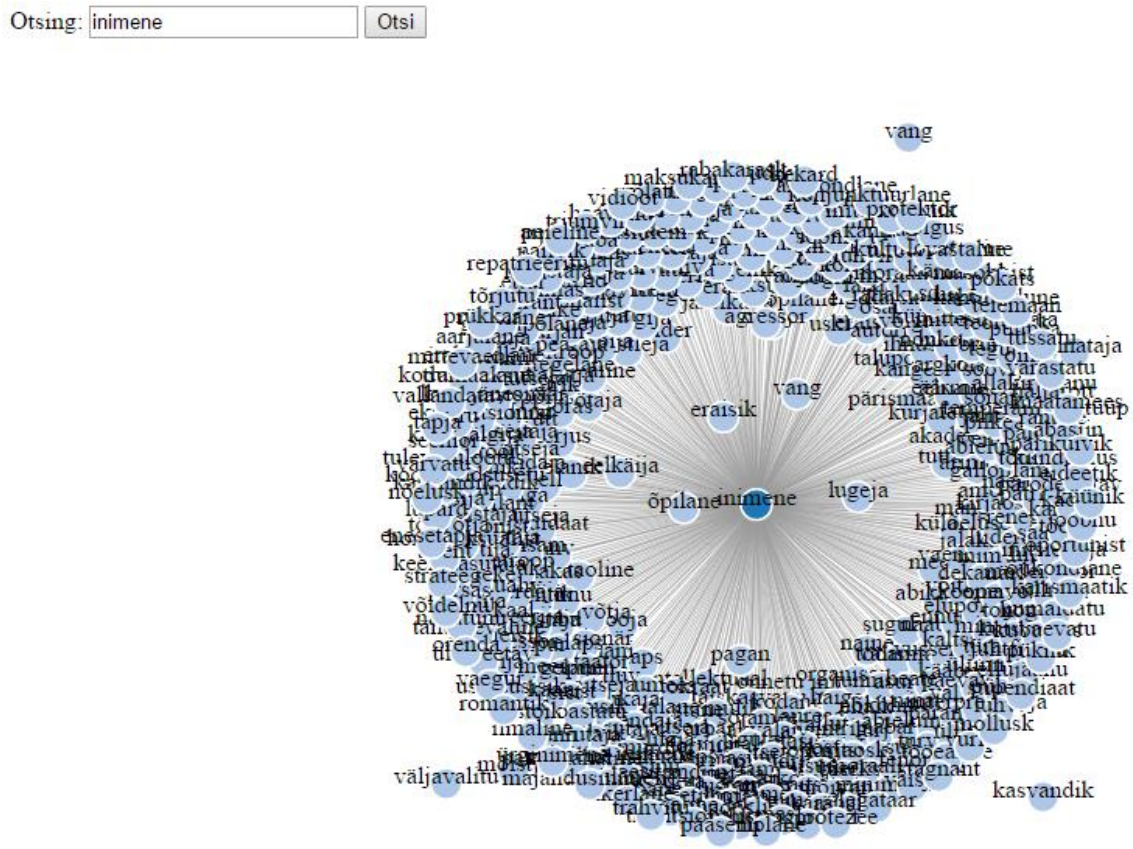
Programmi kujundamiseks kasutati HTML-i ja CSS-i. HTML-iga luuakse dokumendi elementide struktuur ning CSS-iga stiliseeritakse neid elemente. CSS-iga määratakse sõnaseoste ühenduste värvid ning stiilid, samuti teksti suurused ja kujundus.

Fail „index.html“ sisaldab programmi struktuuri: otsingukasti, otsitavat sõna koos definitsiooni ja liigiga ning kolme tulpa („Legend“, graaf ja „Ajalugu“). „Legend“ ja „Ajalugu“ on fikseeritud laiusega, graafi ja sõna *container*’i laius sõltub lehe laiusest.

[11] https://bl.ocks.org/mbostock/4062045 (11.05.2017)

4.3 Teised katsetused

Esialgses versioonis asetsesid sõnad mullide peal, mille värv sõltus seose liigist. See variant langes ära, kuna, nagu on näha joonisel 15, on infot äärmiselt raske hoomata. Pärast mullide eemaldamist muutus pilt arusaadavamaks ning seoseid tähistasid eri värvi jooned.



Joonis 15 Esimesed katsetused

Samuti esialgselt sisaldas programm vaid nelja erinevat suhet – süno- ja antonüümiat, hüpero- ja hüponüümiat. Osadel sõnadel on aga palju ka teisi mõisteid, mistõttu kaasati kõik *wordnet*'i seosed.

4.4 Edasiarendus

Programm täidab püstitatud eesmärki, kuid sellel on ka palju edasiarendusvõimalusi.

Üks võimalus on ühendada antud visualiseerimisprogramm teiste keelte *wordnet*'idega, et kasutaja saaks vabalt liikuda eri keelte vahel. See on suureks abiks nii keele õppimisel kui ka näiteks erinevates keeltes eri kirjatükkide kirjutamisel. Samuti on võimalik lisada sõnasageduse näitajaid näiteks tekstikorpustes ning hüperonüümiapuu tulemusi, st mitmendal astmel otsitud sõna hüperonüümiapuus on.

Võimalik on arendada ka programmi nii, et vale sõna sisestamisel pakutakse välja kõige lähem olemasolev sõna. Programmi on võimalik muuta nii, et topeltklikkides sõnaga seotud mõistel ilmuvad omakorda selle sõna seosed kõik samale ekraanile, eelmist graafi kustutamata.

Kokkuvõte

Käesoleva bakalaureusetöö eesmärk oli visualiseerida Eesti Wordneti sõnastik ja see eesmärk sai täidetud.

Teoreetilises osas antakse ülevaade *wordnet*-tüüpi sõnastiku ajaloost ja ülesehitusest, tutvustatakse semantilisi suhteid. Lisaks on võimalik tutvuda kahe erineva *wordnet*'i visualiseerimisprogrammiga.

Praktilise osa eesmärk oli visualiseerida Eesti Wordnet. Programmeerimiskeelega Python töödeldi esialgset EstWN-i faili nii, et alles jääks vaid antud tööks vajalik info. Visualiseerimine põhineb demol „Force-Directed Graph", mille koodi modifitseeriti antud töö eesmärgile vastavalt. Praktilises osas kirjeldatakse programmi üldiselt, selle algoritmi, tehnilist poolt ja edasiarendamise võimalusi.

Visualiseerimine aitab Eesti Wordneti tegijatel avastastada vigu – juba programmi algusjärgus avastati näiteks, et sõnal „kass" on märgitud antonüümiks sõna „tsiibetkaslane", mis on viga ja mille EstWN-i koostajad ära parandasid. Samuti aitab see palju arusaadavamalt esitada antud sõnastikku.

Viidatud kirjandus

Bostock, M. For Protovis Users. 2012. https://d3js.org/ (14.04.2017)

Eesti Keele Instituut. Eesti keele seletav sõnaraamat. http://eki.ee/dict/ekss/

Eesti Keele Instituut. Vene-eesti sõnaraamat. www.eki.ee/dict/ves/

Eesti Keeleressursside Keskus. https://keeleressursid.ee/et (01.05.2017)

Eesti Keeletehnoloogia. Riiklik programm "Eesti keeletehnoloogia 2011-2017". https://www.keeletehnoloogia.ee/et/EKT2011-2017-programm-uuendet.pdf/view (14.04.2017)

Eesti Keeletehnoloogia. Eesti Wordnet'i täiendamine. 2014. Riiklik programm "Eesti keeletehnoloogia 2011-2017". http://vana.keeletehnoloogia.ee/ekt-projektid/eesti-wordneti-taiendamine (17.03.2017)

Eesti Keeletehnoloogia. Eesti Wordneti täiendamine 2. 2017. Riiklik programm "Eesti keeletehnoloogia 2011-2017". https://www.keeletehnoloogia.ee/et/ekt-projektid/eesti-wordneti-taiendamine-2/eesti-wordneti-taiendamine-2 (01.05.2017)

Erelt, M., Erelt, T. ja Ross, K. Eesti keele käsiraamat. 1997. https://www.eki.ee/books/ekk09/index.php?p=6&p1=4

Fellbaum, C. ja Tengi, R. WordNet, A lexical database for English. 2017. Allikas: Princeton University: http://wordnetweb.princeton.edu/ (03.03.2017)

Few, S. Data Visualization for Human Perception. The Interaction Design Foundation. 2015.

Heinmets, K. *Sõnade leksikaalsed suhted*. 2017. http://leksikaalsedsuhted.weebly.com/leksikaalsed-suhted.html (11.05.2017)

Johannsen, A. ja Seaton, M. 2017. WordTies: http://wordties.cst.dk/wordties-estwn/ (11.05.2017)

Langemets, M. ja Kallas, J. Sõnaraamatud arvutis ehk elektrooniline leksikograafia. *Oma Keel*, 2014, nr 1, lk 35-42.

Loos, E. E., Day Jr, D. H., Jordan, P. C. ja Wingate, J. D. What is lexical database? Dallas: SIL International DigitalResources. 2003

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. Introduction to WordNet: An On-line Lexical Database. *International Journal of Lexicography*(3), 1990, pp 235-244. Allikas: WordNet.

Orav, H., Kerner, K. ja Parm, S. Eesti Wordneti hetkeseisust. *Keel ja Kirjandus*, 2011, nr 2, lk 96−106.

Orav, H., Zupping, S. ja Vare, K. Leksikosemantiliste suhete hägusus Eesti WordNetis. *Emakeele Seltsi aastaraamat*, Tallinn: Teaduste Akadeemia Kirjastus, 2014, nr 60, lk 171-194.

Pajusalu, R. Sõna ja tähendus. Tallinn: Eesti Keele Sihtasutus. 2007

Parm, S. ja Orav, H. Üle 65 500 eesti mõistega arvutisõnastik. *Sirp, 2014*.

Payne, T. Summary of Semantic Roles and Grammatical Relations. 2007. http://pages.uoregon.edu/tpayne/EG595/HO-Srs-and-GRs.pdf (15.04.2017)

Pedersen, B., Linden, K., Vider, K., Forsberg, M., Kahusk, N., Niemi, Nygaard, L., Seaton, M.; Orav, H.; Borin, L.; Voionmaa, K.; Nisbeth, N. J., Rögnvaldsson, E. Nordic and Baltic wordnets aligned and compared through WordTies. Linköping: Linköping University Electronic Press. 2013

Prabhat, S. Difference Between Dictionary and Thesaurus. http://www.differencebetween.net/language/difference-between-dictionary-and-thesaurus/ (14.04.2017)

Szymański, J. ja Chodor, M. WordNet Editor. http://wordventure.eti.pg.gda.pl/ (03.05.2017)

Tartu Ülikooli arvutilingvistika uurimisrühm. Eesti Wordneti koduleheküIg: cl.ut.ee/ressursid/teksaurus/index.php?lang=et (03.03.2017)

W3Techs. Usage of JavaScript for websites. https://w3techs.com/technologies/details/cp-javascript/all/all (07. 05 2017)

Wiley, J. Experiments on Semantic Memory and Language Comprehension. Pittsburgh: Carnegie Mellon University. 1972.

Zapata Becerra, A. A handbook of general and applied linguistics. Andes: Universidad de Los Andes. 2000.

Lisa

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, **Elina Pankrašin**,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose **Wordnet-tüüpi sõnastiku visualiseerimine**, mille juhendajateks on Heili Orav ja Sven Aller,
   1.1 reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2 üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus, **11.05.2017**