TARTU ÜLIKOOL

HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND

EESTI JA ÜLDKEELETEADUSE INSTITUUT

Sandra Eiche

# Eri meetodeid *wordnet*-tüüpi sõnastiku kontrolliks Eesti Wordneti näitel

Magistritöö

Juhendajad: Heili Orav, Ph.D

Sven Aller, MSc

Tartu 2020

# Sisukord

# Sissejuhatus

Eesti Wordnet on 1998. aastast koostatav tesaurus, millesse koondatud mõisted on seotud semantiliste suhete kaudu, et edasi anda mõistetevahelist struktuuri nii, nagu see võiks esineda keelekasutaja meeles. *Wordnet*'id on keeletehnoloogias olulised ressursid, mida saab kasutada näiteks sõnatähenduste ühestamisel. Eesti Wordnetti koostatakse käsitsi, mis tingib sisu detailsuse. Sellegipoolest kaasneb inimfaktoriga tahtmatult vigu ning kuna seni on Eesti Wordnetis keskendutud peaasjalikult tesauruse täiendamisele, püüab siinne töö leida eri viise, kuidas *wordnet*-tüüpi sõnastike korrektsuse kontrollimiseks automaatselt tuvastada potentsiaalseid veakohti.

*Wordnet*'i hierarhiates leiduvad vead võivad seisneda nii ebakorrektses struktuuris kui ka sisus. Siinne töö kontrollib, kas *wordnet*'is leidub juhtumeid, mil üks sünohulk on teisega seotud mitme suhte kaudu või on sünohulk paigutatud liiga üldise hüperonüümi alluvaks, kuigi leidub ka täpsem ülemmõiste. Samuti on hüpoteesiks, et distributiivset semantikat saab kasutada tuvastamaks erinevaid semantilisi vigu nagu *wordnet*'ist puuduvad tähendused ja suhted või väärad hüperonüümid. Eesmärk on koostada programmid, mis aitaks seesuguseid vigu leida, ning katsetada, kas valitud meetodid on vigade leidmiseks õigustatud. Nii saab ka edaspidi töös käsitletud meetodite tingimustest lähtuvalt kontrollida *wordnet*'i korrektsust. Kahe esimese veatüübi kontrollimisel kasutatakse vaid Eesti Wordneti sisest informatsiooni, kolmanda puhul ka distributiivse semantika mudelit.

Eesmärgi täitmiseks on materjaliks valitud Eesti kirjakeele sagedussõnastik, kuna on oluline, et keeles enimkasutatavad sõnad oleksid tesauruses hästi kirjeldatud. Samuti on sagedaste sõnade kasutamine oluline distributiivsel semantikal põhineva katse juures.

Töö koosneb kaheksast peatükist, millest esimesed kolm kirjeldavad teoreetilisi aluseid, millele katsed toetuvad. Neljas peatükk tutvustab samas valdkonnas varem tehtud uurimusi. Sellele järgnev peatükk kirjeldab katsetesse valitud materjali ning valiku põhjuseid. Viimased kolm peatükki kirjeldavad eri tüüpi vigade leidmiseks tehtud katsete meetodeid ning tulemusi.

# 1. *Wordnet*

*Wordnet*'id on leksikaal-semantilised andmebaasid, mida saab kasutada ühe vahendina loomuliku keele töötlemisel (ingl *natural language processing,* NLP). Esimene *wordnet* – ingliskeelne Princeton WordNet[1] – loodi aga hoopis teisel eesmärgil. 1980ndatel tekkis Princetoni ülikoolis töötaval psühholingvist George A. Milleril huvi, kas leksikaalseid üksusi saab struktureerida hierarhilise võrguna nii, nagu need võiksid esineda inimese teadvuses. Sellest mõttest kantuna valmis esimene omataoliste semantiliste võrkude seast, mis on oma masinloetava struktuuri tõttu nüüd pigem arvutilingvistide ja keeletehnoloogide kui psühholingvistide huviorbiidis. (Fellbaum 2010: 231) Praeguseks on *wordnet*'e koostatud väga paljudele erinevatele keeltele[2].

*Wordnet*'i põhiüksuseks on sünonüümsetest sõnadest koosnevad **sünohulgad**, mis on omavahel ühendatud semantiliste suhete kaudu, moodustades nii mõistete võrgustiku. Sama sõnavorm võib korraga kuuluda mitmesse sünohulka, sel juhul on tegemist polüseemse sõnaga. Sünonüümiale lisaks on *wordnet*'is põhiliseks suhteks hüperonüümia ehk alammõiste ja ülemmõiste vaheline suhe (nt *maja* on teatud liiki *hoone*). Nii moodustuvad hüperonüümiapuud, kus ühest juursünohulgast kasvab välja mitmetasandiline struktuur, mille iga tasand on eelmisest spetsiifilisem.

Nimisõnade seisukohast on samuti oluline meronüümia ehk osa-terviku suhe (nt *mootor* on osa *mootorpaadist*). Sõnaliigiti on keskmine puu sügavus erinev ja ka olulisemad suhted võivad erineda (Fellbaum 2010: 233–234). *Wordnet*'i semantilised suhted ei pruugi eri keelte tesaurustes olla samad, nt EuroWordNeti[3] projektis osalevad *wordnet*'id sisaldasid võimalust ühendada omavahel süntagmaatiliselt seotud substantiive ja verbe (Fellbaum 2010: 238).

Keeltele on *wordnet*'e loodud erinevaid strateegiaid kasutades, näiteks käsitsi, poolautomaatselt või automaatselt, samuti on andmebaasi tegemist alustatud nii puhtalt lehelt kui ka Princetoni WordNeti tõlkides (Orav jt 2014: 172). NLP-ülesannetest on

[1] Princeton WordNet: https://wordnet.princeton.edu/
[2] Vt Global WordNetAssociation http://globalwordnet.org/resources/wordnets-in-the-world/
[3] EuroWordNet: http://projects.illc.uva.nl/EuroWordNet/

*wordnet*'i kõige rohkem püütud rakendada sõnatähenduste ühestamisel (ingl *word sense disambiguation*). Samuti on seda kasutatud ontoloogiate ja teadmusbaaside, näiteks YAGO (Suchanek jt 2007) loomisel. *Wordnet*'idele annab lisaväärtust ka keeltevaheline seostatus nt keeltevahelise indeksi InterLingualIndex (ILI) kaudu (Fellbaum 2010: 239).

## 1.1. Eesti Wordnet

Eesti Wordnet[4] (edaspidi EstWN) on eesti keele jaoks koostatav tesaurus, mille loomisega tehti algust 1998. aastal EuroWordNeti projekti raames. EstWNi koostatakse peamiselt käsitsi Princetoni WordNeti ja EuroWordNeti põhimõtetest lähtudes. (Orav jt 2014: 173–174) Princetoni WordNetiga on see seotud ILI kaudu. EstWNis on kasutusel järgmised suhted:

- Hüperonüümia ja hüponüümia:
  - *has_hyperonym* – on teatul viisil (verbi korral), teatud liiki (nimisõna korral)
  - *has_hyponym* – on üks viis (verbi), on üks liik (nimisõna)
  - *has_xpos_hyperonym* – on teatul viisil, teatud liiki (eri sõnaliikide vahel)
  - *has_xpos_hyponym* – on üks viis, on üks liik (eri sõnaliikide vahel)
- Osalus- ja rollisuhted:
  - *role* – mängib rolli
  - *involved* – kaasneb
  - *role_agent* – mängib tegijana rolli
  - *involved_agent* – kaasneb tegija
  - *role_patient* – mängib tegevusobjektina rolli
  - *involved_patient* – kaasneb tegevusobjekt
  - *role_instrument* – mängib vahendina rolli
  - *involved_instrument* – kaasneb vahend
  - *role_location* – mängib kohana rolli
  - *involved_location* – kaasneb koht
  - *role_target_direction* – mängib tegevuse sihtkohana rolli
  - *involved_target_direction* - kaasneb tegevuse sihtkoht

[4] Eesti Wordnet: https://www.cl.ut.ee/ressursid/teksaurus/

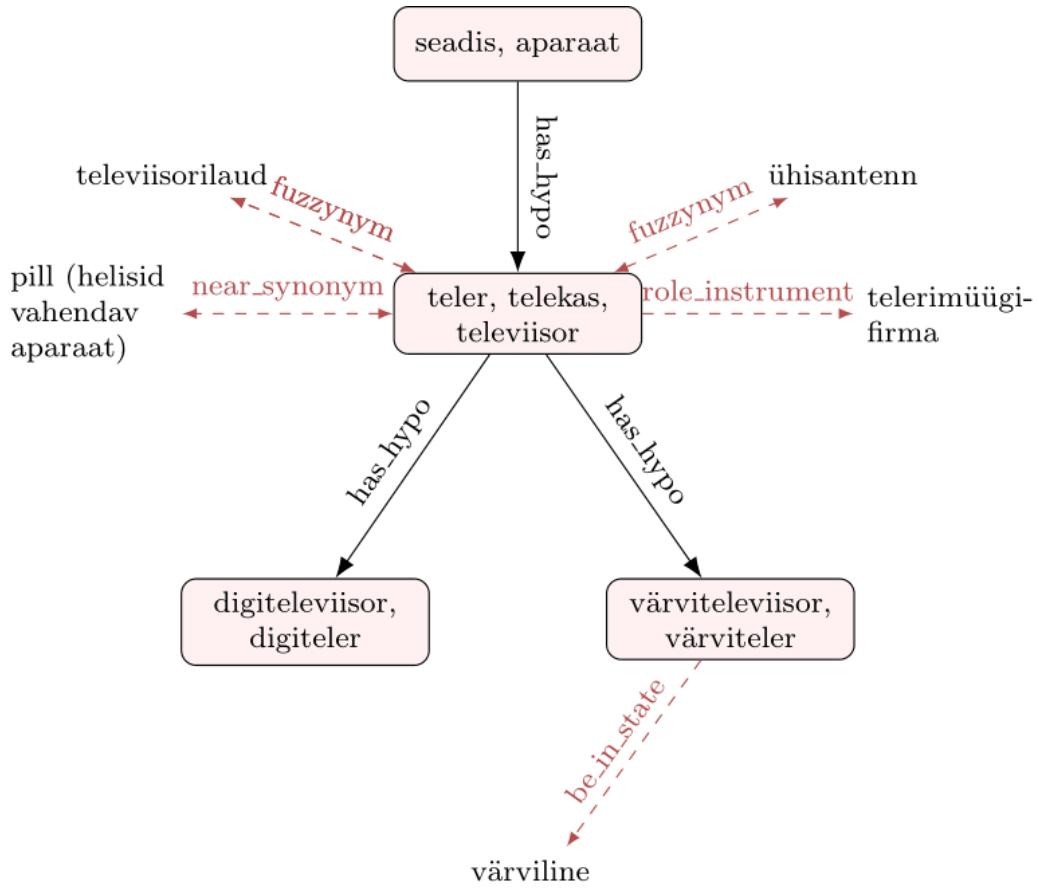
- Meronüümia:
    - *has_holonym* – on osa
    - *has_meronym* – osa on
    - *has_holo_member* – on liige
    - *has_mero_member* – liige on
    - *has_holo_madeof* – on osa materjalist
    - *has_mero_madeof* – üks osa (materjalist) on
    - *has_holo_part* – on osa
    - *has_mero_part* – osa on
    - *has_holo_portion* – on annus
    - *has_mero_portion* – üks annus on
    - *has_holo_location* – on osa kohast
    - *has_mero_location* – üks osa (kohast) on
- Lähisünonüümia:
    - *near_synonym* – peaaegu samatähenduslik on
    - *xpos_near_synonym* – peaaegu samatähenduslik on (erinevate sõnaliikide vahel)
- Antonüümia:
    - *antonym* – vastand on
    - *near_antonym* – peaaegu vastand on
    - *xpos_near_antonym* – peaaegu vastand on (erinevate sõnaliikide puhul)
- Põhjussuhe:
    - *is_caused_by* – on põhjustatud
    - *causes* – põhjustab
- Osasündmus:
    - *has_subevent* – osasündmus on
    - *is_subevent_of* – on osasündmus
- Esindaja:
    - *has_instance* – esindaja on (mõiste seostamiseks pärisnimega)
    - *belongs_to_class* – kuulub klassi

- Hägussuhe:

  *fuzzynym* – on kuidagi seotud

  *xpos_fuzzynym* – on hägusalt seotud (erinevate sõnaliikide vahel)

Joonisel 1 on kujutatud väljavõte EstWNi sünohulka *televiisor.n.01* sisaldavast puust, kus alam- ja ülemmõisteid seovad mustad nooled ning muid suhteid suunatud punktiirjooned.




*Joonis 1. Sünohulga televiisor.n.01 suhted*

Nagu on näha ka jooniselt 1, kasutatakse EstWNis ka suhteid, mida eeskujuks võetud Princetoni WordNetis ei esine. Nende hulka kuuluvad näiteks erinevat tüüpi rollisuhted (näited a ja b), kaasnemissuhted ja hägussuhe (näide c).

(a) *õppija* mängib patsiendina rolli mõiste juures *õppima*,

(b) *haamer* mängib instrumendi rolli mõiste juures *haamerdama*,

(c) *arst* on hägussuhtes mõistega *stetoskoop*. (Orav jt 2014: 184–186)

EstWN on kasutust leidnud näiteks alusleksikonina sõnatähenduste ühestamisel (Zirk 2013), meelestatuse analüüsiks vajaliku ressursi loomisel (Jaanimäe 2018).

## 1.2. Veatüübid *wordnet*'is ja *wordnet*'i valideerimine

*Wordnet*'ides leiduvaid vigu saab laias laastus jagada kolme tüübi vahel: formaalsed, semantilised ja strukturaalsed vead (Piasecki jt 2013: 258). Lohk (2015) on jaganud kõikvõimalikud *wordnet*'is esineda võivad vead nende kolme rühma vahel, formaalsete vigade asemel eelistab ta kasutada terminit „süntaktilised vead".

Süntaktiliste vigade alla kuuluvad Lohki liigituse järgi eksimused failistruktuuris või vigases sisendis, nt kirjavead, puuduv või dubleeritud metainfo jms. Semantilised vead on eksimused sünohulkades ja nendevahelistes suhetes, muu hulgas puuduvad või ebasobivad suhted ja vale definitsioon ehk gloss. (Lohk 2015: 59–60) Piasecki jt (2013) artiklis arvavad autorid, et semantiliste vigade hulka peaksid kuuluma ka morfoloogilise analüsaatori tuvastatud trükivead lemmades või definitsioonides, kuna otsustamaks, kas tegu on tegeliku trükivea või analüsaatorile lihtsalt tundmatu sõnaga, on vaja inimest. Strukturaalsete vigade hulka kuuluvad *wordnet*'i loogilise struktuuri rikkumised, mida saab parandada teades vaid suhtetüüpi ning nagu formaalsete/süntaktiliste vigade jaoks, pole ka nende parandamiseks vaja semantilist analüüsi. (Piasecki jt 2013: 258) Näiteks on strukturaalsed vead ühepoolsed sümmeetrilised suhted või ühel sünohulgal mitme vahetu hüperonüümi esinemine (Piasecki jt 2013: 265). Mõlemat tüüpi vigu saab üldiselt hõlpsalt ennetada, kasutades spetsiaalseid *wordnet*'idele mõeldud toimeteid (nt WordnetLoom[5]), mis võimaldavad *wordnet*'i hierarhiat graafiliselt esitada ning suhteid vaid mõne hiirevajutusega muuta. Toimetis saab näiteks suhete nimetused eeldefineerida ja see hoiab ära trükivead, mida inimene käsitsi sisestades võiks teha. (Piasecki jt 2013: 258–259)

*Wordnet*'i suhete õigsust (vead suhetes võivad kuuluda nii semantiliste kui ka strukturaalsete vigade alla) saab kontrollida mitmel moel. Üks võimalus on käsitada *wordnet*'i kui ontoloogiat ning leida olemasolevas hierarhias vastuolusid, kasutades ontoloogiatele välja töötatud meetodeid (vt ptk 4). Samuti saab suhteid hinnata ühisloome

[5] WordnetLoom: http://nlp.pwr.wroc.pl/en/tools-and-resources/tools/wordnetloom

(ingl *crowdsourcing*) abil. Näiteks Soome FinnWordNeti koostamisel valideeriti sealseid sünonüümiasuhteid, paludes tesauruse kasutajatel hinnata soomekeelse sõna sobivust viiepalliskaalal sõnaliigi, hüperonüümi ja ingliskeelse glossi põhjal. (Lindén, Niemi 2014: 196). Sama võimalust on ära kasutatud ka *wordnet*'i ja Wikipedia vahele loodud ühenduste hindamiseks (Szymanski, Boiński 2019). Sagedasti leitakse mõistetevahelistes suhetes vigu reegleid moodustades ja järgides. Hüperonüümiasuhet on kontrollitud näiteks vaadates sõnastiku definitsioone või leides, kas liitsõna põhisõna kuulub mõnda tesauruses juba olemasolevasse sünohulka ja kas need on üksteisega hüperonüümia kaudu seotud, nt *paperwork – work.* Samuti on suhte valideerimiseks otsitud korpusest hüperonüümi-hüponüümi paari esinemist hüperonüümiat implitseeriva mustri kaudu (nt $w_1$ *such as* $w_2$) (Nadig jt 2008: 4–7). Selle reegli abil leitakse väga spetsiifilisi vigu, sest vaadeldakse vaid liitsõnu. *Wordnet*'i struktuuri hindamiseks on võimalik kasutada ka testmustreid ehk hierarhiates leiduvaid alamstruktuure, mis võivad viidata võimalikule veale (vt ptk 4). Testmustrite järjepidev rakendamine võib tesauruse kvaliteeti oluliselt parandada.

# 2. Distributiivse semantika mudelid

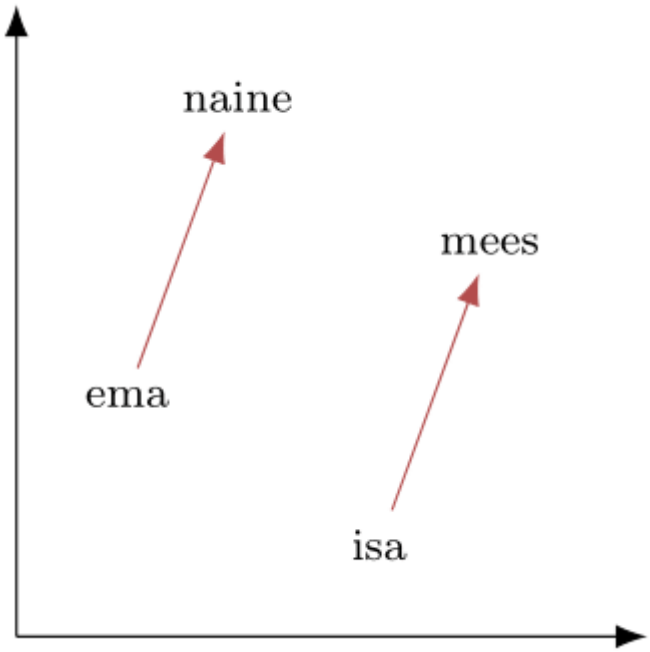
Distributiivne semantika on uurimisvaldkond, mille eesmärk on kirjeldada keeleüksuste tähendusi nende esinemise jaotuse järgi korpustes. Sellega tehti algust 1950ndatel, kui leiti, et sõnu ümbritseva konteksti järgi on võimalik teha järeldusi sõnade sarnasuse kohta. Näiteks kirjeldab Harris (1954), et elemendid (nt sõna, foneem) A ja B on komplementaarsed juhul, kui ühe kontekstis esineb alati X, mida teise kontekstis kunagi pole. Sõnade A ja B kontekstide erinevus või sarnasus viitab analoogsele suhtele nende sõnade tähenduste vahel. Juhul, kui kaht sõna ümbritsev kontekst pole kunagi sama, võib eeldada, et need kuuluvad eri grammatilistesse klassidesse. Sõnu, mis esinevad enamasti samasuguses kontekstis välja arvatud juhtudel, kui ühes lauses on esindatud korraga mõlemad sõnad, nimetab Harris sünonüümseteks. (Harris 1954: 157) Viimast tähelepanekut väljendab distributiivse semantika hüpotees, mille järgi ümbritseb tähenduselt sarnaseid sõnu sarnane kontekst.

Arvutilingvistika jaoks on see hüpotees oluline, kuna korpuste olemasolul on väga lihtne sõnade jaotuste järgi leida nende tähendusi esitavaid representatsioone (ingl *embeddings*) ja peaaegu kõigis korpusepõhistes lähenemistes semantikale on oluline osa konteksti teadmisel. Distributiivse semantika põhimõttele tuginedes on loodud hulk distributiivse semantika mudeleid, mida on kutsutud ka vektorruumi ja semantilise ruumi mudeliteks. (Bruni jt 2014: 1) Tähendusi modelleerivast mudelist võiks olla kasu NLP-ülesannetes nagu küsimustele vastamine ja kokkuvõtete tegemine. (Jurafsky, Martin 2019: 95) Lisaks sellele, et distributiivse semantika mudelit saab mitmekülgselt kasutada, on seda mugav ka treenida. Selle tegemiseks pole vaja spetsiaalset märgendamist ja õppimisprotsess on automaatne ning juhendamata (Jurafsky, Martin 2019: 100).

Mudelis on iga sõna esindatud *n*-mõõtmelise vektorina, kus *n* tähistab tunnuste ehk dimensioonide arvu. Tunnuseks võivad olla nt teised sõnad ja vektor mõõdab sõnade koosesinemist konteksti põhjal. Mudelit võib seega ette kujutada kui *n*-mõõtmelisse ruumi paigutatud sõnade kogumit. Mida lähemal on sõnad vektorruumis üksteisele, seda sarnasemad on nad omavahel. (Aedmaa 2016: 8–9) Vektoritega saab teha lihtsaid matemaatilisi tehteid, näiteks hinnata kahe sõna vahelist sarnasust koosinussarnasuse (ingl *cosine similarity*) kaudu, mis osutab, kui suur on nurk kahe sõna vektorite vahel ehk

kui sarnased on sõnade tähendused (kattuvate vektorite puhul on koosinussarnasus 1, vastassuunaliste puhul -1) (Mikolov jt 2013b:748–749). Samuti saab leida analoogiaid sõnapaaride vahel. Sõnavektorid võivad esindada nii sõna süntaktilisi kui ka semantilisi omadusi, st sõnade analoogia võib seisneda nii vormis kui ka tähenduses. (Mikolov jt 2013a)

Joonisel 2 on kujutatud sõnade *mees*, *isa*, *naine* ja *ema* paigutust lihtsuse mõttes kahemõõtmelises ruumis. Sellelt on näha, et sõnade *mees* ja *isa* vaheline suhe ruumis on analoogne sõnade *naine* ja *ema* vahelise suhtega. Seega tehes sõnavektoritega tehte „isa“ **–** „mees“ + „naine“, võiks tulemus olla lähedane sõna *ema* vektoriga. Süntaktiline analoogia võiks välja tulla tehtest „söön“ + „laulsin“ **–** „laulan“, mille tulemus peaks olema vektorruumis sõna *sõin* läheduses.




*Joonis 2. Sõnade "mees", "naine", "ema" ja "isa" paigutus kahemõõtmelises ruumis*

# 3. *Word2vec*

Üks populaarsemaid ja sagedamini kasutatavaid tööriistu distributiivse semantika mudelite loomiseks on neurovõrgul põhinev vabavaraline *word2vec* (edaspidi W2V). See võimaldab luua mudeleid kahel eri moel.

CBOW (ingl *continuous-bag-of-words*) mudelit treenitakse korpusmaterjalidel nii, et see ennustaks sihtsõna seda ümbritseva *n*-suurusega konteksti kuuluvate sõnade (nt kaks sõna enne ja pärast sihtsõna) vektorite põhjal. Selle mudeli kasuks on treenimisel otsustatud väiksema ajakulu tõttu (nt Aedmaa 2016, Ebert jt 2016).

*Skip*-grammi (ingl *skip-gram*) mudelit treenitakse sihtsõna vektori põhjal ennustama seda ümbritsevat konteksti. (Mikolov jt 2013b: 4). *Skip*-grammi mudeli treenimine on küll ajakulukam, kuid just semantiliste ülesannete lahendamisel on see andnud tunduvalt paremaid tulemusi kui CBOW mudel. Mudelite treenimisel tuleb valida ka dimensionaalsus. Üldiselt on teada, et suurem dimensioonide arv kirjeldab sõnadevahelisi seoseid täpsemalt. Samas on arvu tõstmine treenimise ajakulu ja mudeli kvaliteedi paranemist arvesse võttes kasulik vaid teatud piirini: kui tõsta dimensioonide arvu, kuid mitte korpusandmete hulka ja vastupidi, paranevad tulemused üsna vähe. Treeningandmete suurendamine aga tõstab ka treenimise ajalist keerukust. (Mikolov jt 2013b: 6–8)

Eesti keele koondkorpuse[6] põhjal on eesti keele jaoks CBOW ja *skip*-grammi tüüpi *word2vec*'i mudelid juba valmis treenitud ja kättesaadavaks tehtud[7]. Valikus on lemmatiseeritud korpuse ja sõnavormide peal treenitud nii 100- kui ka 200-dimensioonilised mudelid.

*Word2vec*'i mudeleid kasutades tuleb silmas pidada, et ühele sõnavormile vastab üks vektor, seega ei eristata mudelis homonüümiat ega polüseemiat ja nt sõna *kurk* vektoresitus on saadud kontekstide põhjal, milles *kurk* võis esineda nii köögivilja kui ka kehaosa tähenduses. Samas võib olenevalt mudeli loomise meetodist tekkida ka vastupidine mure: kuna igale sõnavormile vastab erinev vektor, võib üks mõiste olla

[6] Eesti keele koondkorpus: https://www.cl.ut.ee/korpused/segakorpus
[7] *Word2vec*'i mudelid: https://github.com/estnltk/word2vec-models

esindatud paljude erinevate vektoritega. Seega juhul kui treenimiseks on kasutatud lemmatiseerimata andmeid, on sõnadel *naine* ja *naise* mudelis kummalgi oma vektoresitus. Samuti on distributiivse semantika mudelite ühe puudusena välja toodud, et kuna mudelite aluseks olevast korpusmaterjalist ei pruugi tihti välja tulla inimesele iseenesestmõistetavaid teadmisi, nt deskriptiivset infot nagu banaan on kollane või hiirel on pea, siis pole sellised lihtsad assotsiatsioonid mudelis esindatud (Bruni 2014: 1).

# 4. Varasemad tööd

Eesti Wordneti koostamise põhirõhk on olnud peamiselt selle käsitsi või poolautomaatsel täiendamisel (Orav jt 2014: 174). Spetsiaalselt tesauruses juba olemasoleva info korrastamiseks ja kontrollimiseks on seni kasutatud hierarhiapuudes potentsiaalseid vigu ja ebakõlasid sisaldavate kohtade leidmist testmustrite abil. Näiteks on nende kaudu tuvastatud selliseid strukturaalseid vigu, kus ühel sünohulgal on kahe erineva haru kaudu ühendus sama ülemmõistega (Lohk 2015: 78). Samuti on uuritud vertikaalset polüseemiat, mille puhul (päritud) hüperonüümi ja selle hüponüümi tähistab sama sõna. Neil puhkudel võib juhtuda, et tähendusi on üleeristatud ning puud üle vaadates saab sellesse parandusi teha (Lohk jt 2019: 396–398).

Guarino ja Welty (2004) töötasid taksonoomiate ontoloogilise korrektsuse kontrollimiseks välja OntoCleani metodoloogia, mis seisneb selles, et taksonoomias esinevatele mõistetele määratakse kontseptuaalsed metatunnused (*rigidity, identity, unity* and *dependence*) ning seejärel kontrollitakse kindlate reeglite järgi, kas ülem- ja alammõiste metatunnused on kooskõlas. Seda metodoloogiat on võimalik rakendada ka *wordnet*'idele, kuid selle kahjuks räägib ressursinõudlikkus, sest käsitsi märgendamine võtab palju aega ning inimtööjõudu.

Nadig jt (2008) kasutasid sõnastikke, et valideerida sünonüüme ja hüperonüüme. *Wordnet*'is vaadeldavale sõnale leiti vaste sõnastikust ja tuvastati, kas selle definitsioonist tuleb välja mõni tesauruses selle sõnaga seotud sünonüüm või hüperonüüm. Võrreldi ka kahe *wordnet*'is sünonüümsena märgitud sõna definitsioonide kokkulangevuse ulatust sõnastikus. (Nadig jt 2008: 3−5) Jaapani ja Princetoni WordNetis puhul püüti sarnastel viisidel avastada sünonüümide valepaigutusi ehk juhtumeid, mille puhul sõna ei vasta oma sünohulga definitsioonile, kuid selleks kasutati tesauruses endas sisalduvat informatsiooni. Iga sünohulga sõna korral eraldati teised sünohulgad, kuhu see sõna veel kuulub. Saadud hulki võrreldi ja leiti nende ühisosa, st vaadati, kas samasse sünohulka kuuluvad sõnad esinevad koos ka mõnes muus sünohulgas. Hüpotees oli, et samasse sünohulka kuuluvad sõnad on ka teistes sünohulkades rohkem koos esindatud kui kaks suvalist sõna. (Hirao jt 2014)

Mitmel moel kasulikuks keeleressursiks on osutunud sõnavektorid. Neid on muu hulgas kasutatud ka seoses *wordnet*'iga. Näites on *wordnet*'i abil hinnatud distributiivsete mudelite headust. Kapočiūtė-Dzikienė ja Damasevicius (2018) valisid leedu keele sõnavektorite kvaliteedi uurimiseks Leedu WordNeti[8], võrreldes erinevaid mudeleid selle järgi, kas mingile sõnale koosinussarnasuse kaudu leitud kümnest kõige sarnasemast sõnast on mõni esindatud ka selle sõna (vähemalt kahest leksikaalsest üksusest koosnevas) sünohulgas.

Sarnasel viisil, kuid vastupidisel eesmärgil, on *wordnet*'i ja *word2vec*'i kasutanud Loukachevitch ja Parkhomenko (2019). Autorid eraldasid kõigepealt vene keele *wordnet*'is RuWordNet[9] mingi mõistega tihedalt seotud sünohulkadesse kuuluvad sõnad. Seejärel leidsid nad *word2vec*'i mudeli abil samale sõnale 20 sarnasemat sõna. Edasi vaatasid nad sünohulkadest eraldatud sõnade ja *word2vec*'ist saadud sõnahulga ühisosa, eeldades, et kui ühisosa puudub või on väga väike, on põhjust kahtlustada probleemi, mida tuleks lähemalt uurida, nt sõna tähendusmuutust, mida *wordnet* veel ei kajasta. (Loukachevitch, Parkhomenko 2019: 17–18).

[8] Leedu WordNet: https://korpus.sk/ltskwn_en.html
[9] RuWordNeti koduleht: http://www.ruwordnet.ru/en

# 5. Materjal

Töös on andmete töötlemiseks kasutusel Python 3.5 ja Pythoni EstNLTK[10] 1.4 teek, kuna see on mõeldud eesti keele jaoks ning sisaldab ka EstWNi kasutamiseks mõeldud moodulit. *Word2vec*'i mudeli töötlemiseks kasutan teeki Gensim[11], mis võimaldab lihtsasti leida sõnale etteantud arvu lähimaid sõnu, kasutades selleks koosinussarnasust.

Töö põhilisteks ressurssideks, mille põhjal potentsiaalseid vigu ja puudujääke leida, on eesti kirjakeele sagedussõnastik[12] ja EstWN. Sagedussõnastikus leiduva 10 000 lemma hulgast eraldasin vaatlemiseks sõnad, mis on esindatud ka EstWNis, kokku 8548 sõna (edaspidi sagedussõnastik). Selles EstWNi versioonis sisaldas vähemalt ühte sagedussõnastiku sõna kokku 13 125 erinevat sünohulka. Töös on sünohulkade märkimiseks kasutatud sünohulga nimetust vastavalt EstNLTK mooduli *wordnet* sünohulga atribuudile *name*, mis tähistab sünohulka kujul lemma.sõnaliik.tähenduse_id. Keeles sagedate sõnade kasutamine kindlustab, et EstWN sisaldab suuremat osa neist ning et korpuses on piisavalt materjali, et vektorid kirjeldaksid sõnu võimalikult hästi. Mitme veatüübi puhul on oluliseks mõisteks **lähimad suhted**, mille all mõistan sünohulki, mis on vaatluse all oleva sünohulgaga vahetult (st ühe serva kaudu) seotud mõne leksikosemantilise suhte kaudu. Nende hulka kuuluvad ka hüperonüümia- ja meronüümiasuhtes olevad sünohulgad, mille puhul vaatlen ka päritud suhteid. Iga katse juures on täpsustatud, mis sügavusel olevaid hüpero-, hüpo-, holo- ja meronüüme, lugedes vaatluse all olevast sünohulgast alates, lähimate suhete alla valitakse. Lähimate suhete ning nende lemmade leidmiseks on töös loodud programmides kasutusel moodul *closest_relations.py*[13].

Teine oluline töövahend on *word2vec*'i mudel, mida kasutan, et leida sõnu, mis sõnavektorite järgi võiksid olla mingi sihtsõnaga seotud, kuid millega EstWNis seost pole. Kasutan lemmatiseeritud Eesti keele koondkorpuse põhjal treenitud 200-dimensioonilist *skip*-grammi mudelit. Selle kasuks räägib *skip*-grammi parem

[10] EstNLTK 1.4: https://github.com/estnltk/estnltk/tree/1.4.1.1
[11] Gensim: https://radimrehurek.com/gensim/
[12] Eesti kirjakeele sagedussõnastik: https://www.cl.ut.ee/ressursid/sagedused/
[13] https://github.com/eisandra/estwn-validation/tree/master/programs

semantiliste ülesannete lahendamise võimekus, samuti võiks 200-dimensiooniline mudel sõnadevahelisi seoseid täpsemalt kirjeldada kui 100-dimensiooniline. Arvestades seda, et eesti keel on morfoloogiliselt keeruline ning mõne sõna algvormi võib esineda korpuses liiga vähe, et selle põhjal mudel häid tulemusi annaks, on mõistlik kasutada lemmatiseeritud korpusel põhinevat mudelit. Sel moel koondub ühe mõiste kohta käiv info algvormi vektori alla, hoolimata sellest, mis vormis algtekstis sõnad olid. Algvormis kokku langevate sõnade puhul tuleb siiski tähele panna, et kui sõnavormidena võisid olla tähendused kuidagi mudelis eristatavad (nt vaadates sõnavorme *tamme* ja *tammi*), siis lemmatiseeritud mudelis eristatavus kaob. Lemmatiseerimise kasu morfoloogiliselt rikaste keelte mudelite loomisel isegi vähese dimensioonide arvu ja treeningmaterjali puhul on oma uurimuses esile tõstnud ka Ebert jt (2016). Samuti on lemmadele taandamine kasulik, kuna EstWNis on esindatud vaid algvormid.

Kuna *word2vec*'i mudel võib sisaldada palju mittekirjakeelseid lemmasid või *wordnet*'i seisukohalt ebavajalikke numbreid, sümboleid jms, siis nende välja filtreerimiseks aktsepteerin vaid lemmasid, mis on eraldatud Eesti Wordnet 2.3.2 XML-failist, ning Eesti kirjakeele sagedussõnastiku seda osa, mis EstWNist veel puudus.

# 6. Eesti Wordneti sünohulkade suhetest puuduvate lemmade leidmine *word2vec*'i abil

*Wordnet*'i kvaliteedi parandamiseks on võimalik kasutada distributiivse semantika mudeleid. Loukachevitch ja Parkhomenko (2019) näitasid RuWordNeti põhjal, kuidas mudelist võivad välja tulla sõna tähendused, mida *wordnet*'is veel pole. Samuti võib mudelist välja tulla sõnu, mis sobiksid asendama mõnd olemasolevat suhet, nt liiga üldist hüperonüümi. (Loukachevitch, Parkhomenko 2019) Peamine põhjus, miks kasutada distributiivse semantika mudelit nagu *word2vec*, seisneb selles, et mudeli andmed põhinevad reaalselt kasutataval keelel. Selle töö üks eesmärk on leida meetod, kuidas eraldada EstWNi sünohulkadesse kuuluvatele sõnadele W2V mudelist potentsiaalselt seotud sõnu, mis sünohulkade suhetest veel välja ei tule.

## 6.1. Meetod

Potentsiaalselt puudulike sünohulkadega sõnade leidmiseks rakendan igale sagedussõnastiku sõnale ehk sihtsõnale järgnevat lihtsustatud algoritmi:

(1) eraldada sihtsõnale kõigist seda sisaldavatest sünohulkadest lähimate suhete lemmad (edaspidi **WNi sõnahulk**), lähimad suhted on olenevalt katsest muutuvad;

(2) leida sihtsõna ja WNi sõnahulga elementide koosinussarnasused üksteisele;

(3) valida leitud sarnasustest vähim e eeldatavasti kõige kaugema suhte koosinussarnasus;

(4) eraldada Gensimi meetodiga *most_similar* vaatlusalusele sõnale koosinussarnasuse järgi kõik sõnad, mille sarnasus on suurem või võrdne kui kaugeima suhte koosinussarnasus. Juhul kui kaugeim suhe ei kuulu W2V 1000 sarnaseima sõna sekka, eraldada 1000 esimest meetodi *most_similar* vastet;

(5) kontrollida, kas saadud sõnad esinevad EstWNis või sagedussõnastikus (edaspidi kirjakeelekontroll), kontrolli läbinud sõnad moodustavad **W2V sõnahulga**;

(6) leida W2V sõnahulga ja WNi sõnahulga vahe ja esitada see kahanevalt koosinussarnasuse järgi.

Ülesande lahendamisel vaatan sihtsõna kõigi sünohulkade lemmasid koos, kuna on teada, et W2V ei arvesta sõna tähenduserinevusi. Näiteks kui sõna *tee* puhul vaadata selle sünohulka *tee.n.05* (jook, mida saadakse (purustatud) teelehtede leotamisel vees) või *vahend.n.01* (mingile eesmärgile rakendatud tegevus), siis võimalike puuduvate suhetena läheksid kirja W2Vs väga lähedase skoori saanud *maantee* ja *kõnnitee*, kuna tee kui joogi või vahendi suhetes need ei esine. See ühtlasi tähendab, et nende sünohulkade W2V ja WNi vahed kattuvad suurel määral. Tegelikult katab sõnad *maantee* ja *kõnnitee* aga ära sünohulk *tee.n.03* (avalikuks kasutamiseks (reisimiseks v. vedudeks) avatud tee), mistõttu nende esinemine teiste sünohulkade W2V ja WNi vahes oleks eksitav.

Kirjeldatud meetod põhineb hüpoteesil, et vähima koosinussarnasusega WNi sõnahulga lemma esindab semantiliselt kaugeimat suhet ning seega ülejäänud WNi sõnahulk (e lähedasemad suhted) võiksid ära katta suurema osa W2V sõnahulga elementidest. Selleks, et leida kõik sõnad, mille koosinussarnasus W2V mudelis on suurem kui lähimate suhete kaugeim koosinussarnasus, tuleks kõigepealt leida iga sihtsõna ja kõigi mudelisse kuuluvate sõnade vaheline sarnasus. Kuna see on tarbetult ajamahukas, piirdun töös W2V mudelist igale sõnale 1000 lähima sõna leidmisega. 1000 sõna võiks sisaldada kõiki lähimaid suhteid, arvestades seda, et keskmiselt on ühe sagedussõnastiku sõnaga seotud lähimate suhete kaudu u 60 lemmat (koos taksonoomiliste õdedega u 150). Seega kui WNi sõnahulga vähim koosinussarnasus on väiksem kui mudeli sarnasusjärjestuse 1000. element, jääb vaatluse alla 1000 sõna, mis läbivad lisaks kirjakeelekontrolli ja moodustavad W2V sõnahulga. Tulemusena saadav W2V ja WNi sõnahulkade vahe peaks sisaldama sõnu, mis on WNist veel puudu.

Statistiliste näitajate leidmiseks kirjutan tulemused sõnade kaupa JSON-faili, mis sisaldab lisaks W2V ja WNi vahele ka W2V järgi kaugeimat ehk 1000. elementi ning iga sõna puhul sünohulkade kaupa kaugeimat lemmat ja seotud lemmasid nende koosinussarnasusega sihtsõna suhtes.

## 6.2. Esimene katse

Esimese katse juures on lähimateks suheteks vahetud suhted, taksonoomilised õed ning hüpero- ja hüponüümid ning holo- ja meronüümid, mis on sihtsõnast kuni kolme serva kaugusel. Kui hüperonüüme on kolm või vähem ehk juurtipp on sihtsõnale väga lähedal, on vaatluse all vaid vahetud hüpero- ja holonüümid, et vältida liigselt üldiseid mõisteid, mis tavaliselt on hüperonüümiapuu ülemises otsas.

Vaadatud 8542 sõnast on vaid 76 puhul WNi sõnahulga vähim koosinussarnasus suurem kui W2V 1000. sõna sarnasus. See tähendab, et W2V ja WNi sõnahulkade vahest saadud sõnahulka piirab enamasti vaid kirjakeelekontroll, mitte suhted. 499 sõna puhul tuli välja, et WNi sõnahulga suurim koosinussarnasus on väiksem kui W2V sarnaste 1000. elemendi oma ehk W2V 1000 lähima elemendi seas ei leidu ühtki WNi seotud suhte lemmat. Nende hulgas on 251 substantiivi, 92 adjektiivi, 89 muutumatut sõna, 48 verbi, 4 pronoomenit, 1 numeraal ja ülejäänud mitmese analüüsiga. Need tuleks käsitsi üle vaadata, et otsustada, kas kahe sõnahulga mittekattuvus tuleb W2V eripärast (nt *kangur* pole W2V lähimaid suhteid vaadates seotud kangakudumisega, vaid tulemus viitab selle sagedasele kasutusele perekonnanimena) või on tegu oluliste seostega, mida tesauruses veel pole.

Sagedussõnastikus leidus kuus sõna, mida W2Vs ei leidunud ja mille kohta seetõttu informatsiooni ei saanud: *jultuma*, *madalale*, *keskpaigas*, *tõemeeli*, *tülpima*, *vastuvool*. Sõna *madalale* puhul on võimalik, et enamasti analüüsiti seda kui vormi adjektiivist *madal* ja muutumatu sõnana harva või üldse mitte, mistõttu lemmatiseeritud korpuse põhjal tehtud mudelist see sõna välja ei tule. *Jultuma* esineb aga enamasti *nud*-kesksõnana. Tulemusteta ehk W2V ja WNi sõnahulkade vaheta jäid esimeses katses 22 sõna. Neist kaks olid adjektiivid, üks nimisõna ning ülejäänud adverbid. Peamiselt on tegu väikeste sünohulkadega, millel polegi suhteid või vähima koosinussarnasusega suhe on siiski väga lähedane. Nt *tahes.b.01* ainus suhe on lähisünonüüm *ükskõik*, mis on ühtlasi W2V järgi kõige sarnasem sõna ja sel juhul WNi ja W2V sõnahulkade vahet polegi. Kõige erandlikumaks osutus nimisõna *latt*, mille kaugeim suhe WNis oli väga väikse sarnasusmõõduga (0.0008), mistõttu oleks võinud eeldada tulemuse saamist. Takistuseks kujunes aga see, et väga suur osa W2V 1000 sarnasemast sõnast olid arvud ega läbinud

sobivuskontrolli. Need, mis aga sobivaks tunnistati (*naelsterling*, *dollar*, *kroon* jt rahaühikud), olid WNi sõnahulgaga kaetud.

Lisas 1 on näha näidet esimese katse väljundit sõnale *tehas*, mis sisaldab selle lähimaid suhteid ja 145 sõnast koosnevat W2V ja WNi sõnahulkade vahet. Tulemust vaadates võib täheldada, et leidub sõnu, mille puhul tuleks suhe sünohulgaga kindlasti luua, nt *tehas*-lõpuga liitsõnad[14] ja kaaluda võiks suhet sõnaga *tsehh*. Liitsõnu lisades tuleks aga arvestada, et EstWNi koostamisel pole võetud eesmärgiks lisada andmebaasi kõiki liitsõnu, kuna nende moodustamine on väga produktiivne ja tihtipeale on tähendus sõnaosistest otseselt tuletatav. Seda, et tulemus sisaldab ka hägusemalt seotud sõnu, millele on suhteid keerulisem määrata, on märgata juba üsna loetelu alguses (mida suurem on koosinussarnasus, seda eespool on sõna tulemuses). Näiteks pole sobivat suhet, mis võiks otseselt siduda tehas.n.01 ja tehas.n.02 sünohulgad sõnadega *silikaatvärv* ja *liimpuitkilp*. Mida kaugemaid sõnu vaadata, seda rohkem tuleb sisse nõrgalt seotud sõnu.

Kuna ilmnes, et katse jaoks valitud muutujad ei piira W2V sõnahulka efektiivselt, peab lähenemist muutma. Üks võimalus on eeldada, et mudel esindab väga hästi ühte (levinumat) sõna mitmest tähendusest, ja valida WNi kaugeimat suhtet esindama mitte kõigi sünohulkade peale väikseim skoor, vaid võtta sünohulkade kaugeimatest kõige suurem skoor. Nt sõna *muusa* sünohulga *muusa.n.01* (loomingulise inspiratsiooni allikas, hrl. meest innustav naine) kõige kaugem suhe on hüperonüümiga *isend* (sarnasus 0,147), teise sünohulga *muusa.n.02* (kunsti või teaduse kaitsejumalanna vanakreeka mütoloogias (EKSS)) kaugeim suhe on hüperonüüm *jumalatar* (sarnasus 0,554). Skooride järgi võiks arvata, et W2V kirjeldab paremini sünohulka *muusa.n.02*, sest selle vähim skoor on teise sünohulgaga võrreldes palju suurem ja selle võiks võtta ka W2V-st sarnasuste eraldamise lähtepunktiks. Vaadates sel moel leitud vähimaid lähisuhete sarnasusi, on keskmine kaugeima suhte koosinussarnasus 0,218. See jääb siiski alla W2V 1000. lähima lemma keskmisele 0,515, millest võib järeldada, et lahendus poleks oluliselt parem. Meetod pole tõhus ka seepärast, et eelise saavad väheste suhetega sünohulgad. *Muusa.n.02* suhetest tulid välja ainult lemmad *jumal* ja *jumalatar*, isegi päritud hüperonüümid ei mahtunud

[14] Suur osa neist puuduvad EstNLTKs sisalduvast wordnetist, kuid on olemas kirjakeelekontrollis kasutatud EstWN 2.3.2 versioonis ning on praegu sõna *tehas* sünohulkade hüponüümideks.

lemmade hulka, kuna need on mitmesõnalised ega saanud seetõttu W2Vs sarnasusmõõtu. Ei saa teada, kas hüperonüüm *üleloomulik olend* on sõnale *muusa* lähedasem kui teise sünohulga ülemmõiste *isend*.

Teine võimalus tulemusi parandada on vaadata põhjalikumalt EstWNi suhteid. Tabelis 1 on suhete kaupa esitatud, mitu korda esines esimeses katses vaatlusalusesse suhtesse kuuluv lemma kaugeima või lähima suhtena. Juhul kui suurim või vähim koosinussarnasus esines mitme suhte all, läksid arvesse kõik vastavad suhted. Võrdluseks on kirjas ka see, kui palju on sagedussõnastiku sõnadel vastava suhtega seotud sünohulki kokku. Tabelist ilmneb, et u 40% kaugematest suhetest on taksonoomilised õed, ning u 23% hüperonüümid, lähima suhtena esinevad need vastavalt umbes kolmandiku ja üle poole võrra vähem. Samuti võib tähele panna, et kuigi ilmselgelt on sõnadel mitu korda rohkem hüponüüme kui hüperonüüme, on kaugeim sõna suurema tõenäosusega hüperonüüm kui hüponüüm. Ootuspäraselt esinevad sünonüümid lähima suhtena 12 korda enam kui kaugeimana, 5 korda enam on lähimaid suhteid ka teisest sõnaliigist lähisünonüümide puhul, samas lähisünonüümia puhul on vahe väike.

| **Suhe** | **Kordi kaugeima suhtena esinenud** | **Kordi lähima suhtena esinenud** | **Suhtega seotud sünohulki kokku** |
|---|---|---|---|
| *taksonoomilised õed* | 3 529 | 2 338 | 445 718 |
| hüperonüümid | 2 005 | 800 | 36 549 |
| *near_synonym* | 864 | 1 081 | 11 213 |
| *near_antonym* | 540 | 237 | 4 149 |
| hüponüümid | 470 | 1 265 | 221 105 |
| sünonüümid (v.a sihtsõna) | 255 | 3 120 | - |
| *xpos_fuzzynym* | 179 | 47 | 1 750 |
| *antonym* | 148 | 240 | 1 531 |
| *state_of* | 114 | 30 | 1 281 |
| *has_xpos_hyperonym* | 104 | 52 | 1 153 |
| *has_xpos_hyponym* | 103 | 63 | 9 033 |
| *fuzzynym* | 88 | 123 | 6 528 |
| *involved_agent* | 47 | 74 | 2 924 |
| *xpos_near_synonym* | 40 | 218 | 1 257 |
| *role* | 26 | 34 | 1 400 |

| | | | |
|---|---|---|---|
| *be_in_state* | 24 | 22 | 552 |
| meronüümid | 16 | 26 | 1 088 |
| *involved* | 16 | 28 | 889 |
| *role_agent* | 15 | 23 | 549 |
| holonüümid | 13 | 24 | 682 |
| *is_caused_by* | 13 | 37 | 790 |
| *causes* | 11 | 27 | 535 |
| *involved_patient* | 8 | 5 | 100 |
| *involved_instrument* | 6 | 10 | 207 |
| *involved_location* | 4 | 2 | 106 |
| *role_patient* | 3 | 0 | 151 |
| *role_instrument* | 2 | 11 | 186 |
| *xpos_near_antonym* | 2 | 2 | 38 |
| *is_subevent_of* | 1 | 3 | 94 |
| *role_location* | 1 | 6 | 289 |
| *has_subevent* | 0 | 8 | 83 |
| *role_target_direction* | 0 | 1 | 37 |
| *involved_target_direction* | 0 | 1 | 12 |
| KOKKU | 8 647 | 9 958 | 751 979 |

*Tabel 1. Sihtsõnadele kaugeimate ja lähimate suhete lemmade ning suhtega seotud sünohulkade jaotus esimeses katses*

## 6.3. Teine ja kolmas katse

Eelmise katse statistikast oli näha, et enamasti ei piiranud valitud lähimad suhted W2V sõnahulka piisavalt. Seetõttu vähendasin teises katses vaadeldavate päritud hüperonüümide hulka, esiteks sihtsõnast kahe serva kaugusele jäävate hüperonüümideni (katse 2.1) ning siis vaid otseste hüperonüümideni (katse 2.2). Kummalgi juhul jäid kaugeima suhte määramisest välja ka taksonoomilised õed, kuid W2V ja WNi sõnahulkade vahe leidmisel läksid need siiski arvesse.

Katse 2.1 puhul muutus piirangu tõttu kaugeima suhte koosinussarnasus 8542 sõnast 4697 ehk enam kui pooltel sõnadel. Sellegipoolest on 98% juhtudest W2V 1000. kaugeim sõna jätkuvalt sarnasem WNi kaugeimast suhtest. Tabel 2 näitab, et hoolimata hüperonüümide vähendamisest, suurenes taksonoomiliste õdede arvelt hüperonüümide hulk kaugeima

suhtena. Hüppeliselt kasvas ka hüponüümide arv kaugeima suhtena. Keskmiselt sai sõna tulemuseks 236 vastet.

| Suhe | Kordi kaugeima suhtena esinenud | Kordi lähima suhtena esinenud |
|---|---|---|
| hüperonüümid | 3 803 | 1 491 |
| hüponüümid | 1 073 | 1 595 |
| *near_synonym* | 921 | 1 137 |
| *near_antonym* | 553 | 254 |
| sünonüümid | 456 | 3 672 |
| *has_xpos_hyponym* | 257 | 73 |
| *xpos_fuzzynym* | 238 | 58 |
| *fuzzynym* | 227 | 184 |
| *has_xpos_hyperonym* | 191 | 55 |
| *antonym* | 157 | 254 |
| *involved_agent* | 131 | 107 |
| *state_of* | 127 | *32* |
| *xpos_near_synonym* | 76 | 233 |
| *be_in_state* | 70 | 29 |
| *involved* | 56 | 35 |
| *role* | 47 | 43 |
| *role_agent* | 43 | 36 |
| holonüümid | 34 | 35 |
| *is_caused_by* | 34 | 50 |
| meronüümid | 29 | 46 |
| *causes* | 28 | 34 |
| *involved_location* | 13 | 2 |
| *involved_instrument* | 12 | 11 |
| *involved_patient* | 12 | 5 |
| *role_instrument* | 8 | 16 |
| *role_location* | 8 | 7 |
| *role_patient* | 7 | 0 |
| *is_subevent_of* | 2 | 3 |
| *involved_target_direction* | 2 | 1 |
| *has_subevent* | 2 | 9 |
| *xpos_near_antonym* | 2 | 2 |

*Tabel 2. Sihtsõnadele kaugeimate ja lähimate suhete jaotus katses 2.1*

Katse 2.2 käigus muutus võrreldes katsega 2.1 veel 2163 sõna puhul kaugeima suhte koosinussarnasus. Hoolimata teise katsega seatud piirangutest, vähenes keskmine tulemuseks saadud sõnade arv vaid mõne sõna võrra ning jätkuvalt on vähemalt 95% sõnadest piiratud vaid kirjakeelsuskontrolliga (vt tabel 3).

Teine katse näitas, et isegi kui kaugeima suhte koosinussarnasus muutub üle pooltel juhtudel, pole ainult hüperonüümide ja taksonoomiliste õdede vähendamisest suurt kasu, sest enamik sõnu jäävad ikka sisuliselt piiramata. Samas on keeruline valida piiramiseks teisi suhteid, sest nagu oli näha tabelitest 1 ja 2, esineb muid suhteid sagedussõnastiku sõnade sünohulkades vähem kui hüperonüüme, taksonoomilisi õdesid ja hüponüüme, mistõttu nende üksikult eemaldamine kaugeimate suhete leidmisest ei tooks kaasa suuri muutusi. Lisaks tuleb suhete piiramisel arvestada, et iga suhte välja arvestamisel kaugeima suhte kandidaatidest, kaovad ka selle suhtega seotud lemmad, mis on sihtsõnale väga sarnased ja seega olulised. Näiteks on katse 2.2 puhul 2364 sõnal koosinussarnasuse järgi nii lähim kui ka kaugeim lemma samast suhtest.

Kuna suhete piiramine valitud moel osutus keerukaks, oli kolmanda katse ülesehitus teistsugune. See põhines ideel, et *wordnet*'i sünohulkade sünonüümid ei pruugi olla tekstis võrdselt sagedased ning sõnade homonüümia ja polüseemia tõttu võib juhtuda, et mõne sõna vektoresitus iseloomustab paremini selle sõna teist tähendusvälja. Näiteks sõna *masin* ühel sünohulgal on rollisuhe sünohulgaga tembutama.v.02, mille lemmade sarnasusskoorid sõnaga *masin* on järgnevad: ('tembutama', 0.2946622967720032), ('käivituma', 0.3437768816947937), ('jukerdama', 0.5410585403442383), ('jupsima', 0.6139044165611267). Nähtavasti esinevad *tembutama* ja *masin* tekstides koos vähem kui *jupsima* ja *masin*. Võib oletada, et sünohulga igas suhtes on vähemalt üks lemma, mis esineb tekstis sagedasti ja on relevantsem kui teised sama suhte lemmad. Seetõttu valisin kolmandas katses kõige kaugemaks suhteks sõna, mis on iga sünohulga suhete parimatest sarnasusskooridest kaugeim. Tulenevalt sellest, et seekord olid vaatluse all iga suhte suurimad koosinussarnasused, mistõttu ei oleks rohkemate suhete uurimine kuigi palju suurendanud tulemuseks saadavate sõnade arvu, läksid katses taaskord arvesse nii taksonoomilised õed kui ka kolme serva kaugusel olevad hüperonüümid. Tabelist 3 on näha, et võrreldes esimese katsega tõusis WNi kaugeima elemendi keskmine

sarnasusskoor enam kui poole võrra, jäädes siiski jätkuvalt alla W2V 1000. elemendi keskmisele, mis oli 0,51. Ligi tuhande võrra vähenes nende sõnade hulk, mille puhul W2V 1000. element on suurema sarnasusskooriga kui WNi kaugeim suhe, kuid siiski moodustavad need u 85% kõigist sõnadest.

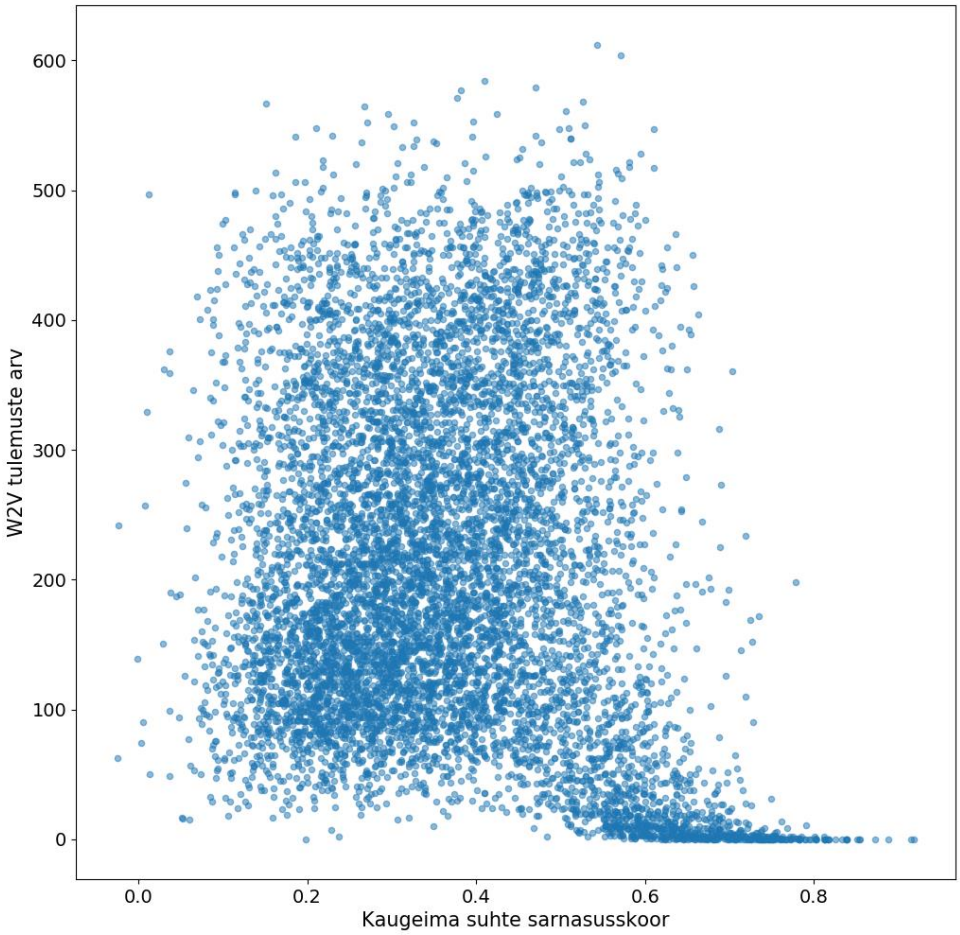
| Katse nr | Keskmine kaugeima suhte koosinussarnasus | Sõnu, millel W2V limiit suurem kui WN-i limiit | Tulemusteta sõnu | Keskmine sõnade arv tulemuses |
|---|---|---|---|---|
| 1 | 0,1517 | 8466 | 22 | 238,2 |
| 2.1 | 0,2106 | 8376 | 30 | 236,2 |
| 2.2 | 0,2417 | 8146 | 61 | 231,2 |
| 3 | 0,3677 | 7248 | 108 | 210,8 |

*Tabel 3. Kolme esimese katse tulemused*

Kolmandas katses said taas eelise väheste suhetega sünohulgad, sest vaadates tulemusteta sõnu, oli lausa 96 sõnal vaid üks sünohulk ja kui välja arvata erandlik sõna *latt*, oli kõigil sõnadel maksimaalselt kolm suhet.

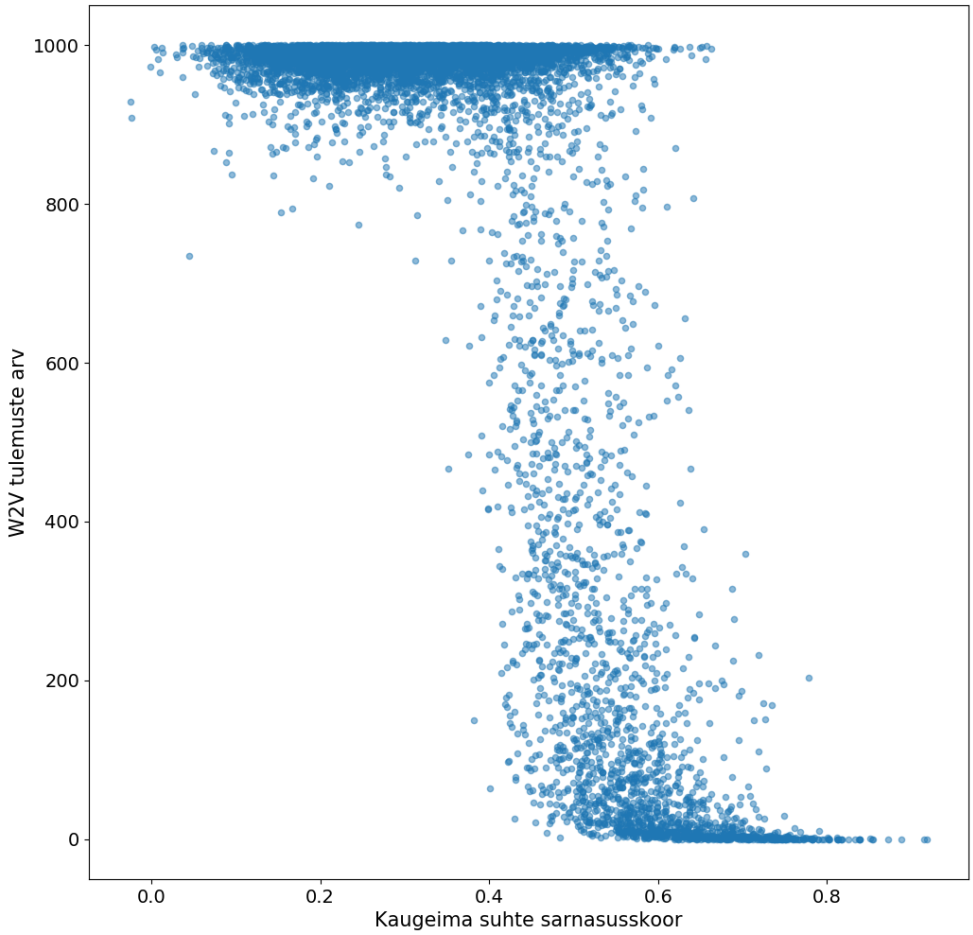
Vaadates kolmanda katse tulemusi jooniselt 3, kus on esitatud tulemusse kuuluvate sõnade arv kaugeima suhte sarnasusskoori järgi, võib täheldada, et juhul kui kaugeima suhte koosinussarnasus on suurem kui 0,6, on tulemuses sõnu vähe, samas sellest väiksemate skooride puhul on tulemused üsna ühtlaselt jaotunud 50–500 sõna vahemikku. Kuna on teada, et suurem osa sõnadest peaksid ilma kirjakeelsuskontrollita saama vasteks 1000 sõna, näitab hajuvusdiagramm selgesti, et nn prahti, mis väljundist filtreeritakse, on eri sõnade puhul väga erinevas koguses, kuid üldjuhul kuulub sellesse vähemalt 500 sõna esialgselt leitud sarnasustest.

Arvestades, et hajuvusdiagrammil joonistub välja saba, kus suhteliselt kõrge sarnasusskooriga kaugeimatel suhetel on vähe tulemusi, võiks arvata, et üldine trend on, et mida väiksem on kaugeima suhte koosinussarnasus, seda rohkem on sõnu W2Vst saadavas tulemuses.




*Joonis 3. Kolmanda katse hajuvusdiagramm tulemuseks saadud sõnade arvu sõltuvusest kaugeima suhte sarnasuse järgi*

Kuna tehtud katsete põhjal on teada, et kaugeima suhte arvestamisel on tulemused ebapraktiliselt suured, võiks trendi uurimiseks muuta lähenemist. Kirjakeelekontroll ei tohiks toimuda alles pärast W2Vst 1000 sarnase sõna leidmist, vaid lähimad sõnad tuleks võtta nende sõnade seast, mille kirjakeelekontroll heaks kiidab. Joonis 4 on hajuvusdiagramm andmetest, mis on saadud kolmanda katse meetodil, kuid mille puhul W2V mudeli sõnavara on piiratud kirjakeelekontrolliks kasutatud sõnadega.




*Joonis 4.Hajuvusdiagramm piiratud sõnastikuga mudelil kolmanda katse meetodigatulemuseks saadud sõnade arvu sõltuvusest kaugeima suhte sarnasuse järgi*

Näha on, et kui sõna kaugeima suhte koosinussarnasus on suurem kui 0,4, on tulemuseks saadud sõnade arv rohkem hajutatud. Tulemuste täpsemaks muutmine piiratud hulgal sõnadel jääb aga selle töö ulatusest välja. Lisas 2 on piiratud mudelil tehtud kolmanda katse väljund, mis näitlikustab seda, kuidas muutub W2V ja WNi sünohulkade vahe, kui eraldada W2V järgi 1000 lähimat suhet vaid varem heakskiidetud sõnade hulgast.

Katsete kokkuvõtteks saab öelda, et kasutatud meetodid ja parameetrid ei pruugi olla efektiivsed distributiivse semantika mudelite kaudu *wordnet*'ist veakohtade leidmiseks soovitud moel. Neist saaks järeldada, et *wordnet*'i sünohulkadest on vaid mõni protsent

hästi kirjeldatud ja ülejäänutest on puudu u 200 seotud lemmat. Seda aga ei saa tulemusi vaadates õigeks pidada, kuna sünohulgad, mis selle hüpoteesi järgi on piisavalt kirjeldatud (e tulemusteta või paari sõnaga tulemuses), on tegelikkuses vastupidiselt väga väheste suhetega. Samuti võib vaid mõnd W2V ja WNi vahet vaadates näha, et suur osa sealsetest sõnadest pole inimese vaatepunktist sihtsõnaga tähenduslikult seotud või on neid väga raske siduda mõne semantilise suhte kaudu. Selles töös vaatasin tulemusi aga pisteliselt, kuna tegemist on mahuka materjaliga ning põhjalikuma analüüsi tegemiseks on vaja, et lingvist vaataks tulemused üle ja annaks neile hinnangu. Katsete väljundid ning programmid on kättesaadaval GitHubis[15].

[15] https://github.com/eisandra/estwn-validation

# 7. Sama sünohulgaga mitme suhte kaudu seotud sünohulkade eraldamine

Teiseks vaatasin sagedussõnastiku sõnade peal veatüüpi, kus sünohulk on teisega seotud mitme suhte kaudu. Tegu on veaga, mida saab avastada suhete struktuuri uurides ja mille tõid oma töös välja ka Piasecki jt (2016: 265). *Wordnet*'i koostajal võib selline juhtum tähelepanuta jääda näiteks mõnd suhet muutes, eriti kui vaatluse all on väga paljude suhetega sünohulk. Samuti võib olla aeganõudev kõigi päritud hüperonüümia- ja meronüümiasuhete läbi vaatamine, et kindlustada ühekordseid suhteid.

## 7.1. Meetod

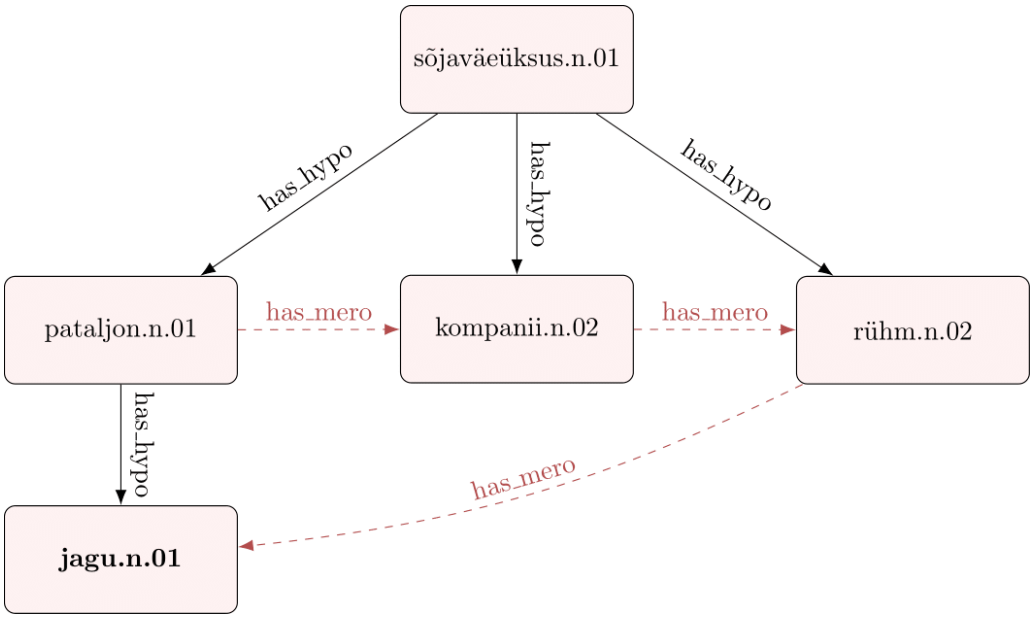
Sama sünohulgaga mitme seose leidmisel on vaatluse all kõik sagedussõnastiku sõnu sisaldavad sünohulgad. Veakohtade leidmisel võetakse arvesse lähimaid suhteid, sealhulgas kogu hierarhia ulatuses hüperonüümiat ja meronüümiat. Lihtsustatult töötab programm järgmiste sammude järgi:

(1) leida sagedussõnastiku sõna igale sünohulgale lähimad suhted, sh sünohulk ise;
(2) leida iga sünohulga kõigi semantilise suhte korral vaatlusaluse suhte sünohulkade hulga ühisosa teiste suhete sünohulkade hulgaga;
(3) kui ühisosa eksisteerib kahepoolsete suhete (hüperonüümid, hüponüümid, hägussuhe, lähisünonüümid, antonüümid) sünohulkade vahel, kontrollida, et sama seos pole varasemaid sõnu vaadates välja tulnud;
(4) kui ühisosa on mitte-kahepoolsete suhete vahel või kolmandas punktis ei ilmnenud varasemat seost, kirjutada väljundisse kirje, mis sisaldab sünohulki, mille vahel on korraga mitu suhet, ja vastavaid suhteid.

Programm salvestab tulemused CSV-vormingusse.

## 7.2. Tulemused

EstWNi põhjal leidsin sagedussõnastiku sõnade seast 129 sellist juhtu, kus sihtsõna mõni sünohulk on seotud ühe või mitme teise sünohulgaga kahe erineva suhte kaudu[16]. Tulemused failis *korduvad_suhted.csv* ja programm *overlapping_relations.py* on leitavad GitHubis[17]. Joonisel 1 on kujutatud sünohulga *jagu.n.01* suhet sünohulgaga *pataljon.n.01*. Need on korraga seotud nii hüperonüümia kui ka meronüümia kaudu. Kui hierarhiat põhjalikumalt vaadata, ilmneb, et mõistlik oleks muuta *jagu.n.01* hüperonüümi, kuna jagu on niisamuti väeüksus nagu selle holonüümidki ja pataljoniga peaks jääma seotuks meronüümiaketi kaudu, st jagu on osa rühmast.




*Joonis 5. Sünohulkade* jagu.n.01 *ja* pataljon.n.01 *vahelised suhted*

Kõige rohkem esines mõne muu suhtega samal ajal koos hüperonüümiat (110 korda), lähisünonüümiat (40), hägussuhet (33) ja meronüümiat (26). See on ootuspärane, kuna need on ka EstWNis sagedasti kasutatavad suhted. Ühel juhul ilmnes ka sünohulga suhe iseendaga. Nimelt sünohulk *side.n.08* (tihedast sidekoest väät, mis ühendab skeleti osi v.

[16] Mõningane kadu võib tulla sellest, kui kahte sünohulka ühendab küll kaks erinevat suhet, kuid ainult sagedussõnastikku mitte kuuluva sõna poolelt (nt kui sagedasel sõnal pole ühtegi be_in_state suhet, kuigi teine sünohulk on sellega seotud suhte state_of kaudu)
[17] https://github.com/eisandra/estwn-validation

elundeid omavahel (EKSS)) on sünohulga *sidekude.n.01* ja selle meronüümi *elastiinkiud.n.01* kaudu iseenda meronüümiks, moodustades nii tsükli.

Suhtepaaridest on kõige sagedamad lähisünonüümia ja hüperonüümia (37), samuti hägussuhe ja hüperonüümia (27). Hüperonüümiasuhe kahe sünohulga vahel juba eeldab seda, et nende vahel on mingi sarnasus (*x* on *y*-i liik), lähisünonüümiat samal ajal nende vahel olla ei tohiks. Näiteks ilmnes tulemustest, et sünohulk *värviline.a.01* on ühtaegu nii hüperonüüm kui ka lähisünonüüm sünohulgale *roosa.a.01*. Hägussuhtele võiks kindlama suhte olemasolul alati eelistada viimast. Hoolimata sellest, et kõiki leitud kattuvaid suhteid võib pidada vigadeks, kuna *wordnet*'i struktuuris seesuguseid kattuvusi olla ei tohiks, pole vea lahendamine automaatselt nii lihtne: kõiki paare on vaja eraldi uurida, mitte eeldada, et spetsiifiline suhe on õigem. Seda ilmestavad sünohulgad *arhitektuur.n.01* ja *ehitussümboolika.n.01*, mille vahel on nii hägus- kui ka hüperonüümiasuhe. Ehitussümboolika kui ehitise osadele sümboolse tähenduse omistamine pole ehituskunsti liik, vaid pigem selle osa, ning vajab seetõttu uut hüperonüümi.

# 8. Hüperonüümia korrastamine taksonoomiliste õdede abil

*Wordnet*'is esineb väga paljude hüponüümidega sünohulki, mis teeb sünohulga haldamise keeruliseks. Näiteks võib koostajal uut mõistet lisades juhtuda, et sobiv hüperonüüm jääb märkamata ning mõiste pannakse üldisema mõiste külge või sobivat hüperonüümi veel pole ning seda hiljem lisades unustatakse alammõisteid ümber paigutada. Vale hüperonüümiasuhe liigitub semantiliste vigade alla ning seetõttu tuleks sellised ebatäpsused puus tuvastada ning parandada. Nadig jt (2008) kasutasid hüperonüümia kontrollimiseks sõnaraamatute kirjeid, liitsõnaosi ja korpusepäringut. Siinses töös püüan kontrollida hüperonüümiat vaid andmebaasis leiduvat informatsiooni kasutades.

## 8.1. Meetod

Lihtsaim viis andmebaasist ülemmõistete leidmiseks on vaadata definitsioone, kuna mõisteid defineeritakse sageli ülemmõistete kaudu. Liitsõna puhul võib ülemmõiste sageli määrata põhisõna järgi. Selleks et mitte piirduda vaid kontrollimisega, kas hüperonüüm sisaldab liitsõna põhisõna või mõnd sõna definitsioonist, võtan potentsiaalsete hüperonüümidena vaatluse alla taksonoomilised õed. Selle suhte kasutamine hüperonüümide eraldamiseks tagab, et vaadeldavad sünohulgad on juba omavahel tähenduslikult seotud, mis ei pruugi aga kehtida juhul, kui otsida potentsiaalseid hüperonüüme kogu *wordnet*'ist või sõnaraamatutest. Siiski eeldab taksonoomiliste õdede kasutamine, et koostaja on paigutanud sünohulga õigesse hüperonüümiapuusse ning võimalikult lähedase ülemmõiste alla. Nii ei saa leida valepaigutusi, mille puhul on sünohulk paigutatud täiesti teise puu alammõisteks.

Selleks, et eraldada mõistetest potentsiaalseid hüperonüüme on vajalikud järgmised sammud:

(1) leida kõik ühe sagedussõnastiku sünohulga alammõistete (edaspidi sihtsünohulk) lemmad;
(2) eraldada iga sihtsünohulga jaoks märksõnad ehk kõik selle sõnaliigiga kattuvad lemmatiseeritud sõnad definitsioonist ja sihtsünohulka moodustavate sünonüümide põhisõnad;

(3) iga sihtsünohulga puhul vaadata, kas mõni märksõnadest kattub taksonoomilise õe lemmaga, kattumise puhul viidata vastavale taksonoomilisele õele kui potentsiaalsele hüperonüümile;
(4) vaadata, kas ühisosasse kuuluvad sünohulgad on esindatud mõne muu lähisuhte kaudu, sh hüpernüümia- ja meronüümiasuhetega kolme serva kauguselt sihtsõnast.

Definitsioonide lemmatiseerimiseks ja tüvede leidmiseks kasutan EstNLTK teeki. Tulemused salvestatakse CSV-formaati, mis sisaldab viit veergu:

1) sagedussõnastiku sõna, mille hüponüüme vaadeldakse;
2) sünohulk, mille taksonoomilistest õdedest leiti sobiv hüperonüümikandidaat;
3) hüperonüümikandidaadid;
4) vaatluse all oleva sünohulga definitsioon, mille põhjal hüperonüüme leiti;
5) saadud soovitused, mis on sihtsõnaga seotud juba muu suhte kaudu.

## 8.2. Tulemused

Programmi tulemusena sai potentsiaalse suhtekandidaadi 802 sõna 3537 sünohulka ehk 27% vaadeldutest. Tulemused failis *seotud_taksonoomilised_oed.csv* ja programm *hypernym_extraction.py* on nähtaval GitHubis[18]. Kõige rohkem said soovitusi sagedussõnastiku sõnade *hing* (132), *haigus* (123) ja *kehaosa* (103) hüponüümid, mida on nende sõnade sünohulkadel ootuspäraselt palju, vastavalt 426, 487 ja 155. Soovitusi sai 2914 nimisõna, 622 verbi ja 2 omadussõna sünohulka. Tulemusi täpsemalt vaadates ilmnes, et lisaks hüperonüümidele (näide a) tuleb definitsioonidest mitmel juhul välja ka hüponüüme (näide b), antonüüme (näide c), meronüümiat (näide d) ja sünonüümiat/lähisünonüümiat (näide e). Juhul kui sõna tähistab osa mingist tervikust või süsteemist, võib definitsioonis esineda ka teisi sama terviku osi, näiteks järgnevuse väljendamiseks (näide f). Viimast kahte seost on hästi näha palju soovitusi saanud *kehaosa* hüponüümide puhul, kuna nende definitsioonides on väga tavaline külgnevate kehaosade või osa-terviku suhte välja toomine.

[18] https://github.com/eisandra/estwn-validation

(a) safiir.n.01 – vääriskivina kasutatav sinine **korund**
(b) rahvajutt.n.01 – folkl pärimuslik proosateos (nt **muinasjutt**, muistend, naljand) (EKSS)
(c) linnatoit.n.01 – linnalik toit (vastandina maa-, **talutoidule**) (EKSS)
(d) kristoloogia.n.01 – Jeesuse Kristuse isikut käsitlev kristliku **teoloogia** osa (EKSS)
(e) kahurituli.n.01 – kõnek. **suurtükituli** (EKSS)
(f) silmahammas.n.01 – lõike- ja **purihammaste** vahel asetsev tugev terava otsaga hammas

Antonüümiasuhe ei avaldu vaid definitsioonides, vaid võib ilmneda ka liitsõnaosiste järgi leitud soovitustes, nt sünohulka subjektiivsus.n.01 kuulub sõna *mitteobjektiivne*, mille põhisõna järgi saab sünohulk suhtesoovituseks vastandi objektiivne.n.01. Otseseid vigu tekitavad eksitavad liitsõnad, nt rahvapärased looma- ja taimeliikide nimetused. Nii saab musträhn.n.01, mis sisaldab rahvapäraseid nimetusi *nõgikikas* ja *metskukk*, soovitusteks nii rähn.n.01 kui ka kukk.n.03 ning võilill.n.01, rahvapäraselt *võikann*, soovituseks kannike.n.01. Samuti tekib probleem, kui mõni taksonoomiline õde, millest soovitusi otsitakse, sisaldab liiga üldist mõistet. Näiteks sisaldab auto.n.01 sõnu *auto*, *käru*, *automobiil*, *sõiduk*, millest viimane esineb sageli liikumisvahendeid tähistavates sõnades liitsõnakomponendina või sisaldub mõiste definitsioonis. Seetõttu on auto.n.01 ebasobivaks hüperonüümisoovituseks nt sünohulkadele tramm.n.01, buss.n.01 ja lumesõiduk.n.01. Sünohulkade definitsioonides võib samuti leida sõnu, mida on keeruline sünohulgaga EstWNi leksikosemantiliste suhete kaudu siduda, nt metanool.n.01 (väga mürgine värvusetu etanooli lõhnaga vedelik) saab suhtesoovituseks etanool.n.01. Definitsioonist on näha, et seal seostatakse kahte ainet lõhna kaudu, seda kirjeldavat suhet aga valikus pole.

Lihtsaim võimalus vaid hüperonüümiasuhte eraldamiseks oleks vaadata definitsioonis nimetavas käändes esinevaid sõnu, sest nagu näidetest näha, on teised suhted enamasti markeeritud käänetega. Samas on ka muud suhted üle vaatamist väärt ning seetõttu pole mõistlik neid lihtsalt kõrvale heita, vaid edaspidi leida mustrid, millega saab tulemused vastavalt suhtele sorteerida.

Verbide puhul on keeruliseks kohaks ühend- ja väljendverbid, kuna lemmatiseerimisel need jagunevad ning edaspidi on vaatluse all vaid verbiosa, nii kaotab sõna aga suure osa oma tähendusest. Nt sõna *muutuma* alluvatest saavad 12 sünohulka suhtesoovituseks jääma.v.04 ja saama.v.07, sest neisse kuuluvad lemmad *saama, jääma, muutuma, minema* on väga mitmetähenduslikud ja esinevad ka ühend- ja väljendverbides. Seepärast on need soovituseks ka sünohulgal katkema.v.03, millesse kuuluvad muu hulgas verbid *otsa saama* ja *järele jääma*.

Soovituse saanud sünohulkadest 349 ehk 9,8% puhul on vähemalt üks soovitus esindatud ka muudes, sh päritud, suhetes. Tabelist 4 on näha, mitu korda on soovitatud taksonoomiline õde olnud vaatlusaluse sünohulgaga suhte kaudu seotud. Kõige enam on esinenud hägussuhet, mida on kasutatud nt järgnevussuhte edasiandmiseks. Holonüümid-meronüümid ning lähisünonüümid ja antonüümid on samuti suhetes hästi esindatud. See, kas mõni suhe on esindatud järjepidevalt korrektselt, et selle olemasolul soovitus eemaldada, vajab edasist semantilist analüüsi. Üldiselt on aga semantilisele veatüübile kohaselt iga soovituse puhul vaja koostaja hinnangut, kas sõna vajab täpsemat hüperonüümi või saab lisada ka mõne muu semantilise suhte.

| **Suhe** | **Soovitatud sünohulga esinemiste kordi lähisuhetes (v.a taksonoomilised õed)** |
|---|---|
| *fuzzynym* | 73 |
| holonüümid | 67 |
| *near_synonym* | 64 |
| *antonym* | 57 |
| *near_antonym* | 49 |
| meronüümid | 18 |
| *causes* | 6 |
| *involved_location* | 5 |
| *involved* | 4 |
| *be_in_state* | 2 |
| *is_caused_by* | 2 |
| *role_target_direction* | 1 |
| *is_subevent_of* | 1 |
| *involved_target_direction* | 1 |
| *state_of* | 1 |

| | |
|---|---|
| hüperonüümid[19] | 1 |
| *role_location* | 1 |

*Tabel 4. Definitsioonidest ja põhisõnadest leitud sünohulkade esinemiste arv eri suhetes*

[19] Saab juhtuda, kuna sünohulgal kevadvesi.n.01 oli kaks hüperonüümi – vesi.n.02 ja sulavesi.n.01, millest esimene oli teise ülemmõiste.

# Kokkuvõte

Wordnet on leksikaal-semantiline andmebaas, mida on loodud paljude eri keelte jaoks, kuna seda saab kasutada keeletehnoloogilistes rakendustes mitmel eesmärgil. Eesti Wordnetti on koostatud üle kümne aasta, suurt rõhku on pööratud andmebaasi täiendamisele. EstWNi on käsitletud ka mõnes töös, mis puudutavad *wordnet*'ist testmustrite leidmist ja seeläbi vigade tuvastamist. Siinse tööga püüdsin anda panuse olemasolevast andmebaasist erinevat tüüpi vigade leidmiseks.

Esimesena käsitlesin distributiivse semantika mudeli kasutamist EstWNist puuduvate mõistete leidmisel. Katsete lõpptulemusena valmis neli JSON-formaadis faili, mis sisaldasid olulisima osana sagedussõnastiku sõnadele *word2vec*'i mudelist leitud sarnaseid sõnu, mida EstWNis veel polnud. Katseteks valitud muutujatega ei saanud kinnitust hüpotees, et kõik sõnad, mis on koosinussarnasuse järgi sihtsõnale sarnasemad kui kaugeim lähisuhe tesauruses, peaksid esindama olulist seost, mis andmebaasist veel puudub. Tulemuseks saadud sõnahulgad olid väga suured ning sisaldasid palju ebasobivaid sõnu. Katsete kasutegurina võib näha hoopis seda, et mõningal määral aitab see tuvastada *wordnet*'is väga väheste sõnadega või halvasti kirjeldatud sõnu selle järgi, et nende suhted pole tulemustes kas üldse esindatud või on vastupidiselt esindatud kõik suhted.

Teiseks vaatlusaluseks veatüübiks oli ühe sünohulga samaaegne seotus teisega mitme suhte kaudu. Sellise vea tuvastamiseks mõeldud programm leidis sagedussõnastiku sõnadele põhinedes 129 vea esinemisjuhtu. Mõnel juhul oli kattuv suhe tekkinud ka päritud suhete kaudu, mis võib ühe juba olemasoleva suhte tähelepanemise koostajale keerulisemaks teha. Kõigile sellistele vigadele *wordnet*'is tuleks leida lahendus.

Kolmandaks ülesandeks oli sünohulgale sobivamate hüperonüümide leidmine taksonoomiliste õdede hulgast sünohulkade definitsioonide ja liitsõnade põhisõnade abil. Kokku leiti hüperonüümiakandidaate 3537 sünohulgale, kuid ilmnes, et lisaks täpsematele hüperonüümidele tuli definitsioonidest palju välja ka teist tüüpi suhteid, nt meronüümiat, hüponüümiat ja antonüümiat. Umbes kümnendikul juhtudest oli mingisugune semantiline side sünohulga ja soovituse vahel olemas.

Töö tulemusena valmisid programmid, mille abil kirjeldatud veatüüpe leida, ning iga veatüübi kohta ka potentsiaalsete veakohtadega nimekirjad, mis EstWNi kvaliteedi parandamiseks tasuks üle vaadata.

Kuigi siinses töös ei andnud distributiivse semantika mudeli kasutamine soovitud tulemusi, võib uurida edaspidi, kas valitud sõnadeni piiratud mudelit kasutades annab vaatlusaluseid suhteid varieerides jõuda väiksemate ja täpsemate tulemusteni. Eelkõige oleks aga vaja inimese hinnangut sellele, milline *word2vec*'i väljund võiks olla *wordnet*'i suhete jaoks sobiv, et parameetreid lihtsamini, ehk isegi automaatselt, valida. Täpsemate hüperonüümide määramise juures võiks kindlasti edasi arendada definitsioonidest eri suhete leidmist. Andmetest joonistusid välja kindlad suhted, mis definitsioonides sisaldusid ning definitsioone uurides oleks võimalik leida igale suhtele omaseid mustreid, mida hiljem saaks kasutada soovituste eraldamiseks suhete kaupa või isegi suhte automaatseks määramiseks.

# Kasutatud allikad

**Aedmaa, Eleri 2016**. Eesti keele ühendverbide kompositsionaalsuse määramine. – Eesti Rakenduslingvistika Ühingu aastaraamat 12, 5–23. http://arhiiv.rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat/article/view/ERYa12.01/360. Vaadatud 17.01.2020

**Bruni, Elia, Nam Khanh Tran, Marco Baroni 2014**. Multimodal Distributional Semantics. – Journal of Artificial Intelligence Research 49, 1–47.

**Ebert, Sebastian, Thomas Müller, Hinrich Schütze 2016**. LAMB: A Good Shepherd of Morphologically Rich Languages. – Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, 742–752.

**Eesti keele koondkorpus;** http://www.cl.ut.ee/korpused/segakorpus/index.php?lang=et. Vaadatud 30.10.2019

**Eesti keele sagedussõnastik;** https://www.cl.ut.ee/ressursid/sagedused/index.php?lang=et. Vaadatud 23.10.2019

**Eesti keele *word2vec*'i mudelid;** https://github.com/estnltk/word2vec-models.

**Eesti Wordnet;** https://www.cl.ut.ee/ressursid/teksaurus/. Vaadatud 10.03.2020

**EuroWordNet;** http://projects.illc.uva.nl/EuroWordNet/. Vaadatud 10.03.2020

**EstNLTK 1.4;** https://github.com/estnltk/estnltk/tree/1.4.1.1. Vaadatud 31.05.2020

**Fellbaum, Christiane 2010**. WordNet. – Theory and Applications of Ontology: Computer Applications. Ed. Roberto Poli, Michael Healy, Achilles Kameas. Springer, 231–243.

**Gensim;** https://radimrehurek.com/gensim/. Vaadatud 31.05.2020

**Global WordNetAssociation;** http://globalwordnet.org/resources/wordnets-in-the-world/. Vaadatud 10.03.2020

**Guarino, Nicola, Cristopher Welty 2004**. An overview of OntoClean – Handbook on Ontologies. Ed. Steffen Staab, Rudi Studer. Berlin, Heidelberg: Springer, 210–220.

**Harris, Zellig S**. **1954**. Distributional Structure. – WORD, 10, 146–162.

**Hirao, Takuya, Takahiko Suzuki, Koki Miyata, Sachio Hirokawa 2014**. Detection of Misplacement of Synonyms in the Japanese WordNet. – Proceedings - 2014 IIAI 3rd International Conference on Advanced Applied Informatics, IIAI-AAI 2014, 31–36.

**Jaanimäe, Gerth 2018**. Eesti Wordnet ja meelestatuse analüüs. Magistritöö. Tartu: Tartu Ülikool.

**Jurafsky, Dan, James H. Martin 2019**. Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/edbook_oct162019.pdf. Vaadatud 19.02.2020

**Kapočiūtė-Dzikienė, Jurgita, Robertas Damasevicius 2018**. Intrinsic Evaluation of Lithuanian Word Embeddings Using WordNet. lk 394–404.

**Leedu WordNet;** https://korpus.sk/ltskwn_en.html. Vaadatud 13.03.2020

**Lindén, Krister, Jyrki Niemi 2014**. Is it possible to create a very large wordnet in 100 days? An evaluation. – Language Resources and Evaluation, kd 48, nr 2, lk 191–201.

**Lohk, Ahti 2015**. A System of Test Patterns to Check and Validate the Semantic Hierarchies of Wordnet-type Dictionaries. Doktoritöö. Tallinn: Tallinna Tehnikaülikool.

**Lohk, Ahti, Heili Orav, Kadri Vare, Francis Bond, Rasmus Vaik 2019**. New Polysemy Structures in Wordnets Induced by Vertical Polysemy. – Proceedings of the Tenth Global Wordnet Conference, lk 394–403.

**Loukachevitch, Natalia, Ekaterina Parkhomenko 2019**. Thesaurus Verifcation Based on Distributional Similarities. – Proceedings of the Tenth Global Wordnet Conference, 16–23.

**Mikolov, Tomas, Kai Chen, Greg Corrado, Jeffrey Dean 2013a**. Efficient Estimation of Word Representations in Vector Space. – Proceedings of Workshop at ICLR.

**Mikolov, Tomas, Wen-tau Yih, Geoffrey Zweig 2013b**. Linguistic Regularities in Continuous Space Word Representations. – Proceedings of the 2013 Conference of the

North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Atlanta, Georgia: Association for Computational Linguistics, 746–751.

**Nadig, Raghuvar, Pushpak Bhattacharyya, J. Ramanand 2008**. Automatic Evaluation of Wordnet Synonyms and Hypernyms. – Proceedings of ICON-2008: 6th International Conference on Natural Language Processing, .

**Orav, Heili, Sirli Zupping, Kadri Vare 2014**. Leksikosemantiliste suhete hägusus Eesti Wordnetis. – Emakeele Seltsi aastaraamat, 60, 171–194.

**Piasecki, Maciej, Łukasz Burdka, Marek Maziarz, Michał Kaliński 2013**. Diagnostic Tools in plWordNet Development Process. – Human Language Technology. Challenges for Computer Science and Linguistics. Ed. Zygmunt Vetulani, Hans Uszkoreit, Marek Kubis. Cham: Springer International Publishing, 255–273.

**Princeton WordNet;** https://wordnet.princeton.edu/. Vaadatud 10.03.2020

**RuWordNet;** http://www.ruwordnet.ru/en. Vaadatud 13.03.2020

**Szymanski, Julian, Tomasz Boiński 2019**. Crowdsourcing-Based Evaluation of Automatic References Between WordNet and Wikipedia. – International Journal of Software Engineering and Knowledge Engineering, 29, 317–344.

**Suchanek, Fabian M., Gjergji Kasneci, Gerhard Weikum 2007**. Yago: a core of semantic knowledge Unifying WordNet and Wikipedia. – WWW ’07.

***Word2vec*’i mudelid eesti keele jaoks;** https://github.com/estnltk/word2vec-models. Vaadatud 30.10.2019

**WordnetLoom;** http://nlp.pwr.wroc.pl/en/tools-and-resources/tools/wordnetloom/. Vaadatud 13.03.2020

**Zirk, Kristi 2013**. Reeglipõhine ühestaja eesti keele jaoks. Bakalaureusetöö. Tartu: Tartu Ülikool.

## Different methods for validating wordnets based on Estonian Wordnet. Summary

Wordnet is a lexical-semantic database that has been made for many languages. It is mainly used for several natural language processing tasks. Estonian Wordnet has been worked on for 15 years and the focus has been on expanding the database by adding new senses and relations. Some research has been done about using test patterns for validatingwordnets that also included validating Estonian Wordnet. This research tried to demonstrate validating Estonian Wordnet by finding various types of errors from the database using words from the Estonian frequency dictionary that were included in Estonian Wordnet as a basis.

Firstly, an attempt was made to use distributional semantics, specifically word2vec models, for finding missing senses or relations for the frequent words. Altogether four JSON-files were created that included sets of words that were missing as relations to frequent words in wordnet, but that were presented as being similar to them according to the word2vec model. Nontheless, the attempt failed to prove the hypothesis that words that are more similar by cosine similarity to the target word than the farthest relation lemma (lemma that is connected to the target word via a relation, but has the smallest cosine similarity out of other related lemmas), are somehow related to the target word and should thus be added as a synset or a relation to the wordnet. Most of the result sets were too big for detailed examination and consisted of many irrelevant words.

Secondly, a program was made to find overlapping relations from the thesauri. It meant finding cases where two synsets were connected by two or more different relations at the same time. 129 such mistakes were found. Sometimes the overlapping was between inherited relations, which makes finding the mistake more complicated for a human contributor.

The third way of validation was finding if a word might have more suitable hypernyms among its co-hyponyms. Definitions and heads of compound words were used for that.Program found possible hypernymy candidates for 3537 synsets, but more detailed look into the data revealed that some of the recommendations were more suitable for other

types of relations, e.g. meronymy, hyponymy and antonymy. In some cases the synset recommendations extracted from definitions were also already present in other relations.

# Lisad

# Lisa 1

Esimese katse väljud sõna *tehas* kohta peatükist „Eesti Wordneti sünohulkade suhetest puuduvate lemmade leidmine word2vec'i abil".

| TEHAS | | | |
|---|---|---|---|
| SÜNO-HULGAD | | | |
| | tehas.n.01 | | |
| | | kaugeim | [0.25364962220191956, ['fuzzynym']] |
| | | suhted | **sünonüümid:** |
| | | | vabrik (0.7608925700187683),<br>tehas (1.0) |
| | | | **fuzzynym:** |
| | | | tööstustoode (0.25364962220191956),<br>töölissöökla (0.3170815706253052),<br>vabrikupood (0.530316948890686) |
| | | | **hüperonüümid**: |
| | | | käitis (0.5532344579696655),<br>tööstusettevõte (0.563244640827179),<br>tootmisettevõte (0.5760365128517151),<br>ettevõte (0.6169883608818054) |
| | | | **hüponüümid**: |
| | | | tuletikuvabrik (0.42956051230430603),<br>elektrijaam (0.4793976843357086),<br>tuumajõujaam (0.518927812576294),<br>hüdroelektrijaam (0.5356215834617615),<br>aatomielektrijaam (0.5455543994903564),<br>tuumaelektrijaam (0.5617613792419434),<br>manufaktuur (0.5625460147857666),<br>tuumajaam (0.5759601593017578),<br>tikuvabrik (0.5804591774940491),<br>laevatehas (0.5907765626907349),<br>keemiatehas (0.681374728679657) |
| | | | **taksonoomilised õed:** |
| | | | saekaater (0.5785998702049255),<br>kombinaat (0.6360291838645935),<br>saeveski (0.6414558291435242) |
| | tehas.n.02 | | |
| | | kaugeim | [0.25364962220191956, ['fuzzynym']] |

| | | |
|---|---|---|
| | suhted | **sünonüümid:** |
| | | vabrikuhoone (0.4990242123603821), tehasehoone (0.5735412240028381), vabrik (0.7608925700187683), tehas (1.0) |
| | | **fuzzynym:** |
| | | tööstustoode (0.25364962220191956) |
| | | **hüponüümid:** |
| | | eksperimentaaltehas (0.5185534954071045), katsetehas (0.6558906435966492) |
| | | **taksonoomilised õed:** |
| | | tootmishoone (0.6184855699539185) |
| | | **hüperonüümid** |
| | | maja (0.2841826379299164), ehitis (0.3032674789428711), hoone (0.33010008931159973), ehitus (0.3740653097629547), tööstusehitis (0.37887367606163025), tööstushoone (0.4216267466545105), ehitatu (0.42216095328330994) |
| W2V TULEMUS | autotehas, ketrusvabrik, tsehh, katsepartii, silikaatvärv, traktoritehas, remonditehas, liimpuitkilp, peakonveier, mootoritehas, pöördahi, klaasitehas, õmblusvabrik, paberivabrik, tsemendivabrik, tootmiskorpus, vagunitehas, kuivsegu, kudumisvabrik, tootmisüksus, gaasianalüsaator, tekstiilivabrik, alumiiniumitehas, puitkiudplaat, jõusöödatehas, ketrustööstus, lehtmetall, vedelkaup, mööblivabrik, metallitehas, metallurgiatehas, kalatöötlemisfirma, leivatehas, masinatehas, tsemenditehas, puitplaat, firma, laevaehitustehas, lennukitehas, õlifirma, hapnikutehas, tseoliit, terasetehas, taaskäivitama, tööstusrobot, bioplast, õlitööstus, tankitehas, sagedusmuundur, sõjatehas, klinker, metallifirma, ehituskeraamika, seavabrik, gaasitehas, keemiafirma, vineerivabrik, turbatehas, kaubandusfirma, plastmassitehas, vineeritehas, kütuseterminal, tootmismaht, kergplokk, konservitehas, katsetootmine, emafirma, elektroonikafirma, tootmine, konteineriterminal, moodulmaja, kärgtellis, võitööstus, kontsern, tellisetehas, laevamootor, autokontsern, tootmiskeskus, juustuvabrik, turbabrikett, tootev, valmistoodang, fassaaditellis, relvatehas, auruturbiin, vaheladu, juustutööstus, turbatööstus, briketitehas, puidutööstus, ravimitehas, metallitöötlemine, põhitootmine, tööstuskompleks, mood, raadiotehas, valukoda, ehitusplokk, tootmishall, autotööstus, jahuveski, seafarm, viinavabrik, kasumets, gaasigeneraator, jalatsivabrik, riidevabrik, turbogeneraator, autoremonditöökoda, tootmisruum, suusavabrik, suur- | |

| | |
|---|---|
| | tööstus, tööstuspark, toorriie, asfaltbetoonitehas, lihatööstus, nioobium, kaevandus, rauamaak, keemiakontsern, tuumareaktor, koostootmisjaam, kasevineer, seinaplokk, abiettevõte, kalatööstus, laevaremont, aastatoodang, elektrituulik, tunnelahi, kergehitis, maasoojuspump, seismapanek, tootmistehnoloogia, müügifirma, piiritusetööstus, la, lade, tantaal, logistikapark, toodang, pliiaku, leivavabrik, tselluloosivabrik, proovipartii |

# Lisa 2

Kirjakeelekontrolli sõnadele piiratud mudelil kolmanda katse meetodil tehtud katse kaugeimad suhted ning W2V ja WNi vahe sõnal *tehas* (ülejäänud väljund sama mis lisas 1).

| TEHAS | |
|---|---|
| tehas.n.01 | **kaugeim**: [0.530316948890686, ['fuzzynym']] |
| tehas.n.02 | **kaugeim**: [0.25364962220191956, ['fuzzynym']] |
| autotehas, ketrusvabrik, tsehh, katsepartii, silikaatvärv, traktoritehas, remonditehas, liimpuitkilp, peakonveier, mootoritehas, pöördahi, klaasitehas, õmblusvabrik, paberivabrik, tsemendivabrik, tootmiskorpus, vagunitehas, kuivsegu, kudumisvabrik, tootmisüksus, gaasianalüsaator, tekstiilivabrik, alumiiniumitehas, tuuma-elektrijaam, puitkiudplaat, jõusöödatehas, ketrustööstus, lehtmetall, vedelkaup, mööblivabrik, metallitehas, metallurgiatehas, kalatöötlemisfirma, leivatehas, masinatehas, tsemenditehas, puitplaat, firma, laevaehitustehas, lennukitehas, õlifirma, hapnikutehas, tseoliit, terasetehas, taaskäivitama\|taaskäivitatud, tööstusrobot, bioplast, õlitööstus, tankitehas, sagedusmuundur, sõjatehas, klinker, metallifirma, ehituskeraamika, seavabrik, gaasitehas, keemiafirma, vineerivabrik, turbatehas, kaubandusfirma, plastmassitehas, vineeritehas, kütuseterminal, tootmismaht, kergplokk, konservitehas, tehas\|tehaste, katsetootmine, emafirma, elektroonikafirma, tootmine, konteineriterminal, moodulmaja, kärgtellis, võitööstus, kontsern, tellisetehas, laevamootor, autokontsern, tootmiskeskus, juustuvabrik, turbabrikett, tootev, valmistoodang, fassaaditellis, relvatehas, auruturbiin, vaheladu, juustutööstus, turbatööstus, briketitehas, puidutööstus, ravimitehas, metallitöötlemine, põhitootmine, tööstuskompleks, moe\|mood, raadiotehas, valukoda, ehitusplokk, tootmishall, autotööstus, jahuveski, seafarm, viinavabrik, kasumets, gaasigeneraator, jalatsivabrik, riidevabrik, turbogeneraator, autoremonditöökoda, tehas\|tehased, tootmisruum, suusavabrik, suurtööstus, tööstuspark, toorriie, asfaltbetoonitehas, lihatööstus, nioobium\|nioobiumi, kaevandus, rauamaag\|rauamaak, keemiakontsern, tuumareaktor, koostootmisjaam, kasevineer, seinaplokk, abiettevõte, kalatööstus, laevaremont, aastatoodang, elektrituulik, tunnelahi, kergehitis, maasoojuspump, seismapanek, tootmistehnoloogia, müügifirma, piiritusetööstus, la\|lade, tantaal, logistikapark, toodang, pliiaku, leivavabrik, tselluloosivabrik, proovipartii, puitlaastplaat, puuvillavabrik, lihakombinaat, mööblifirma, mööblitööstus, elektrivedur, pruulikoda, näp\|näpi\|näpp, metalltarind, timber\|timberi, tsement, autolammutuskoda, taaskäivitama\|taaskäivitanud, kaevanduma, nahavabrik, kalamajand, tsemenditööstus, vabrik\|vabrikute, ekskavaatoritehas, seadmestama, jõutrafo, aatomireaktor, teraviljasalv, keskladu\|kesklagu, kuutoodang, ehitussegu, raudteetehas, kütusehoidla\|kütusehoidlane, keraamikatehas, matkapliit, elektroonikatööstus, valmiskaup, veisefarm, raketitehas, kütusehoidla, pressvorm, | |

kalkunikasvatus, õlletööstus, suhkruvabrik, tubakavabrik, tselluloos, sõidu-auto, puhastusseade, lennukimootor, lennukitööstus, alküüdvärv, katlamaja, kütuseladu|kütuselagu, lammutuskoda, konteinerlaev, tööstus, merekonteiner, marl|marli, tütarfirma, õlletehas, lastija, tetrapakend, reaktor, aidu, frotee, piimatööstus, rikastusvabrik, tootma, edam, muldmetall, terminal, valutsehh, pugin|pugini, null-kasum, gaur|gauri, põlevkiviõli, tooraine, kaalumaja, autotõstuk, laoplats, naftakontsern, viilhall, keskladu, eeltellimus, autotootja, konveiertootmine, ette-võte, pastöör, ehitustarve, harvester|harvesteri, tööriistakauplus, turbiin, puisteaine, katseeksemplar, kasvuturvas, freespink, viljahoidla, tütarettevõte, sigala, autoremondifirma, plastmasspakend, klaasivabrik, masinasaal, naftahoidla, eksporttoodang, tuulejaam, montaažitöö, tuulegeneraator, vedelsõnnik, põhitoodang, puiduhake, jäätisefirma, kuivati, spaa-hotell, olv, turg|turgi|turk, normeerija, tuuleelektrijaam, tolmufilter, remonditöökoda, tööpink, töökoda, reaktiivkütus, õmblustööstus, õlikas, metallitööpink, vorstivabrik, pagaritööstus, väetiseladu, kaablitehas, rehvifirma, kiirrestoran, montaaživaht, tsisternvagun, suurlaut, naftaterminal, farmaatsiatehas, tuuleturbiin, hüdrosõlm, pruunsüsi, aurugeneraator, tarnima|tarninud, lõpptoodang, naftatööstus, kartulihoidla, saepalk, kondiitritööstus, keemiatööstus, lubjaahi, ühisfirma, saneerima|saneeritud, alumiiniumitööstus, konstrueerimisbüroo, ehitusmasin, puistekaup, demonteerima|demonteeritud, depoo, galvaniseerima, ujuvdokk, pagarifirma, joogitööstus, veondusfirma, gaas|gaasi, horm|hormi, kütteladu|küttelagu, tööstuskontsern, ökotoit, sõnnikuhoidla|sõnnikuhoidlane, reisijaam, bioreaktor, autofirma, jaekett, saneerima|saneerinud, kondiitrivabrik, tankla|tanklad, närimiskompvek|närimiskompvekk, täisalv, teenustöö, betoonitehas, raudteemajandus, tarnima, metallitööstus, autokompanii, põletusahi, raadiolokatsioonijaam, villima, mootorrattatehas, tekstiilitööstus, jäätmehoidla, kartulihoidla|kartulihoidlane, põllumajand, naftapuurauk, abitootmine, makaronitööstus, vahtpolüstüreen, laevaehitusfirma, elektri-energia, vedu|veo|veod, ebakvaliteetne|ebakvaliteetse, linatööstus, perefirma, betoonitöö, vagunipark, saematerjal, äriinkubaator, merepuksiir, fansa, suurfirma, õllefirma, gaasimootor, biogaas, rauatööline, külmutusseade, lukufirma, salaviinavabrik, veinitööstus, logistikafirma, suhkrutehas, autoklaav, väetisehoidla, õmblusfirma, firmakauplus, palkmaja, ehituspuit, säästukauplus, puistang, ühisettevõte, ohutusrihm, naftaleiukoht, lubjatehas, liinitööline, mootor|mootori, seeriatoodang, lehtklaas, diopsiid|diopsiidi|diopsiit, falkonett, pealadu, agrofirma, terasprofiil, lihakaup, ühemehefirma, tehnoloog, kahjutustamine, peenpalk, aheraine, õlleköök, varas|varasi, puhasti, veofirma, allettevõte, jää-lõhkuja, emaettevõte, automüügifirma, lukupood, pagaritöökoda, tilge, toor-aine, iste|istme|istmed, põllumajandustööstus, jõetamm, gaasküte, diiselvedur, laohoone, eksimeerlaser, mööbliäri, vankelmootor, portlandtsement, kaevandusasula, investeeringufirma, variaator, gaasiküte, laevaremonditehas, flagman, valvefirma, saneerija, terminal|terminali, sanatooriumihoone, suurettevõte, hävija, drink|drinki, rakis, laienema|laienetud, passistama, tsisternauto, prügila, puurimismasin, õõnespaneel,

pooltoode, greifer, olv|olvi, aerostaat, põlevkivi, maanteeviadukt, kindlustis, gaasimaardla, automaatblokeering, aparaaditehas, tsistern, programmjuhtimine, puistlastilaev, vaheladu|vahelagu, frees, utiliseerima|utiliseeritud, kolkaküla|kolkaküll, tootmistegevus, soojusvahet|soojusvaheti|soojusvahett, ehituskaup, edasimüüja, maaõli, hulgiladu, põlevkiv|põlevkivi, transportöör, kohalevedu, teesiht, puidu|puidud|puit, mööbel|mööbli, hulgimüügifirma, määrdeõli, rääs, kellatehas, lennuterminal, maht-universaal, väetisesegu, masinist, puidusüsi, metsapõlvkond, kompanii, söetööstus, mööblitoodang, jaekaupleja, paakvagun, teaduslinnak, utiliseerima, laborihoone, pisiettevõte, hollender, laevaehitus, rekonstrueerima|rekonstrueerinud, sertifitseerija, tütar-ettevõte, propaan, osa-ühing, betoonkatusekivi, kütuserong, metallivärv, külmhoone, enamus-osalus, allhange, kalatööstusfirma, vahvlitort, komb|kombi|komp, karastusjoogi|karastusjook, tarima|tarimad, loomalaut, aasta-käive, lihamüük, varrastaja, turb|turv|turva, jaemüügifirma, ehitusplaat, kanakasvatus, raketiehitus, valmistaja, biopuhastus, väliskaubandusfirma, jäähokikepp, aidu|aidud, originaalpudel, fosforiit, patenteerima|patenteeritud, kalakonserv, maagaas, messiala, turbamaardla, kunstitoode, elevaator, jäätmehoidla|jäätmehoidlane, vaksalihoone, teeninduspakett, kalandusfirma, trükiplaat, ainumüügiõigus, kaevanduskompanii, laevatööstus, saunsuvila, kompvekivabrik, purusti, paberipuit, betoontala, pesula, monumentaal|monumentaali, kanapihv, jahutusradiaator, smuugeldama|smuugeldatud, metallmööbel, õmblustoode, elektroonikakaup, konservitööstus, allettevõtja, jõelaevandus, fosforiiditehas, restruktureerima|restruktureeritud, pargas, toiduainetetootja, kindlustustöö, muukvõti, kosmoselennuk, pürolüüs, elektri-võrk, põllumajandushoone, kalatöötlemine, gaasitanker, tarbekeemia, veeühing, tööstusharu, utiliseerimine, kaugküttetorustik, sõjatööstus, kahveltõstuk, pumpla, kauba-märk, turbaraba, jaotusseade, peps|pepsi, kummitoode, dte, rõugevaktsiin, vedelgaas, tornikuppel, veetsistern, jalatsitööstus, telefonipoo|telefonipood, kaubakai, traktoripark, silikaadi|silikaat, amfiibauto, kalapood, kütteelement, eriversioon, keskkoolihoone, savikivi, õmblustöökoda, ehitusministeerium, lähi-ümbrus, söekaevandus, elektroonikakompanii, palgiparvetaja, prügi-mägi, toiduainetööstus, hakkpuit, moonaküla, väikemaja, re-eksport, koovit|kooviti, universaalkauplus, protekteerima|protekteeritud, rauakauplus, diopsii|diopsiid, elektritarve, jaotla, söevagun, värvimistöökoda, teraviljakombain, tööline, soojuspump, valmisravim, kütuseladu, vabasadam, mullafrees, rahalehm, isoleerija, kiir-tramm, jäätmejaam, silditama, raudteesõlm, meier|meierei, poehoone, metallkonstruktsioon, isolatsioonimaterjal, nikkel|nikli, veevarustussüsteem, büroo|bürood, meiereihoone, baas|baasi, fosforiidimaardla, läbindaja, metall|metalli, käitama|käitatud, laadimistöö, turuhall, tärklisevabrik, põrnika|põrnikas, kontsern|kontserni, aiand, riisiõlu, tankla|tanklate, tuumalaev, osonaator, betoonisegu, polsterdaja, pontoonsild, stop|stopi|stopp, lüpsikarjalaut, nullkasum, angroo, müügi-ja, soomussõiduk, mööblikauplus, raadioaparatuur, agregaat, vabamajandustsoon, kõrgtehnoloogiafirma, autokraana, villima|villitud, esinduspood, biokütus,

lorup|lorupi|lorupp, alakoormatus, meierei, ratsionaliseerima|ratsionaliseerinud, saneerima, müügiturg, purunenu, punkerdaja, õllepruulimine, transpordifirma, arapp, teng|tengi|tenk, toore, kaubaladu, juurde-ehitus, toorkangas, seisanud|seiskama, kaubandushoone, pakendama, rostvärk, tulives|tulivesi, esinduskauplus, kalevivabrik, gaasihoidla, juveelifirma, makaroniroog, haljastu|haljastus, kümnekordistama, reisiterminal, varuja, seinakattematerjal, kosmoseeksperiment, reisivagun, pepperon|pepperoni, kaabel|kaabli, sisseveolitsents, romuauto, gaur, suurtootja, väiketööstus, individuaalõmblus, limusiinifirma, diiselauto, üheinimesetuba, rida-elamu, takso|taksod, päikese-energia, katsesõit, väikefirma, jaam, bakaalkaup, reservtoide, sanitaarseade, soolakaevandus, atsetüleen, ümarpuit, ehitustrust, raudbetoonkarkass, ehitusfirma, masinatööstus, ainu-omanik, lauavabrik, kaevandaja, gaasiauto, ehitustöö|ehitustööd, artell, teraviljatööstus, avatäide, kombain, keemiakauplus, inglistina, stividor|stividori, veski|veskite, kaevandama|kaevandanud, pihkuma, villimine, raketikütus, toiduõli, õlle|õlu, magneesiumkarbonaad|magneesiumkarbonaat, rauakaup, mustang|mustangi, vasekaevandus, piimavedu, seeriatoode, säästupood, autokaal, plasttoode, vahemahuti, nikkel|nikkeli, õllekoda, sadamaladu, kaubahoov, suitsugaas, abiehitis, distribuutor, angaar, kaubakala, kastell, koloniaalkauplus, tööriistapood, puidutöökoda, jahutussüsteem, tootmisühistu, konservtoit, tootmistööline, vaivundament, kommunaalelamu, käidelnud|käitlema, vara|varad|varade, releekaitse, volta, töölisasula, ladu|lagu, põletusaine, piim|piima, veepaak, võistluskelk, valveinsener, karkasshoone, rootormootor, nikkel-metallhüdriidaku, peaaktsionär, soomustama, sprot|sproti|sprotid|sprott, nahatoode, mahuti, liimpuit, naftaplatvorm, jõujaam, orienteerima|orienteerinud, konstruktsioon|konstruktsiooni, piiritusekanister, terasetootja, toodetud|tootma, lammutustöökoda, kontorihoone, hõljuklaev, subsideerima|subsideerinud, parkla|parklad, fotoäri, külmutuskamber, kelt|kelti, õhu-sõiduk, pumbaruum, nioobium, raudbetoonrajatis, virves, vahendajafirma, aiand|aiandi|aiant, läbivedu, toidupoekett, kergkruus, hoburaudtee, sadamatehas, autokumm, tankla, käitluskulu, tagavaratee, kallurauto, superviisor|superviisori, võidusõidurada, konservikarp, veosekäive, taaserastama, haruldane|haruldas|haruldaste, torus|torusi, väärindama, ruberoidkatus, steel, müügikõlblik, taimekaitseprits, kokkuvedu, meierei|meiereide, olmehoone, haisutama, lendtuhk, tööstuskeskus, laevaremondifirma, tootja, vahelduvvoolumootor, keemiakaup, kinnisvara-arendaja, krupp|kruppi, metsavedu, spordimess, profülaktoorium, maa-apteek, loomakasvatusfarm, autotöökoda, sertifitseerima, välisfirma, keldrikauplus, seeriatootmine, külmlaut, hõbetamine, nahaparkimine, vaarikakasvatus, kvass, gutta, merepaat, erimajanduspiirkond, tuulekanal, betoontee, suigu|suik, sport-auto, lisaaktsia, vesk|veski|veskid, rahvasuusk, sealaut, omakse|omaksed, paene|paeste, üürisuvila, autokaup, lumekoristusmasin, võistlev|võistleva, kütteladu, eravaldaja, kultuuri-asutus, kontorimööbel, kümnekordistama|kümnekordistanud, laevastik|laevastiku, gaasitoodang, tisleritöökoda, jook|jooki|jookide, leid|leidi|leit,

tulekahi|tulekahju, ladestama|ladestatud, lõh-keaine, teehöövel, sadam, amortiseerima|amortiseerinud, desintegraator, höövelmaterjal, ümarmaterjal, tankija, müügiterminal, kaubaplats, silohoidla, suitsukast, aastaeel-arve, turvas, pottahi, krematoorium, tangitud|tankima, pond|pondi|pont, kaarhall, silokombain, alpinaarium, vehklemismask, laevatama, äri-plaan, kammivabrik, noorkarjalaut, tuulemootor, kaevandama|kaevandatud, masstootmine, eksportima, tüüp-projekt, depoo|depood, rõivatööstus, veekanister, keskkonna-mõju, kaasaskantav|kaasaskantava, teraskatus, parfümeeriakauplus, seebivabrik, tuulepark, ag, vastumanööver, viljakuivati, korrodeeruma|korrodeerunud, kaubanduskett, laomees, tankur|tankurite, stividor, sovhoosikeskus, parandustöökoda, ehituskeemia, kõrgahi, sprott, tollivaba, terasekompanii, üldkogum|üldkogumine, käsipesu, sisseseade, kloorimine, nuluõli, seeb|seebi|seep, sõi-duauto, esilinastama, trammipark, tsentraalkatlamaja, ladu, mobiiltelefonifirma, kahtlane|kahtlas, algkoolihoone, pugin, generaator, profiilplekk, võistlusjaht, erimajandustsoon, külakauplus, kosmosetehnika, oblastikeskus, kokkuostupunkt, sõjaväehoone, nüüdisajastama, pisifirma, tarbepaber, esmatarbekauplus, juurviljahoidla, tridens|tridensi|tridenss, neljakordistama, sadamarajoon, reisikorraldusfirma, tulepaak, metallitöökoda

## Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Sandra Eiche,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Eri meetodeid wordnet-tüüpi sõnastiku kontrolliks Eesti Wordneti näitel“, mille juhendajad on Heili Orav ja Sven Aller, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

*Sandra Eiche*

***02.06.2020***