TARTU ÜLIKOOL
Arvutiteaduse instituut
Informaatika õppekava

Kristo Markov

Arvuti sõnastike seletuste automaatne genereerimine ja kontrollimine Eesti Wordneti näitel

Bakalaureusetöö (9 EAP)

Juhendajad: Heili Orav,
Indrek Jentson

Tartu 2022

Arvuti sõnastike seletuste automaatne genereerimine ja kontrollimine Eesti Wordneti näitel

Lühikokkuvõte:

Käesoleva töö eesmärk oli automaatselt genereerida seletusi Eesti Wordneti neile mõistetele, millel seletused puuduvad. Töö teoreetilises osas kirjeldatakse arvuti leksikonide ajalugu ning antakse ülevaade wordnet-tüüpi leksikonide põhimõtetest. Täpsemalt kirjeldatakse Eesti Wordneti loomist, sisu ja erinevaid probleeme. Lisaks seletatakse, kuidas koostada ja formaliseerida sõnaseletust.

Töö praktilise osa tulemusena valmis programm, mis genereerib neljal erineval meetodil seletusi mõistetele, millel Eesti Wordnetis need puuduvad. Praktilise osa käigus jõuti järelduseni, et seletusi on võimalik genereerida, kuid kõik genereeritud seletused tuleb üle kontrollida, sest ükski meetod ei andnud 100% korrektset seletust. Kokku genereeriti 11 075 sõnaseletust 18 731 puuduvast seletusest. Kõige rohkem seletusi (5469 seletust ehk ligi 50% seletustest) genereeriti unikaalse sünohulga liikme põhjal. Kõige parema täpsusega töötasid sarnasuse (91% seletustest sobis) ning unikaalse sünohulga liikme meetodid (84% seletustest sobis).

Võtmesõnad:

arvuti sõnastik, Eesti Wordnet, sõnaseletuste automaatne genereerimine

CERCS: P170 Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)

Programme for automatic generation and verification of computer dictionary definitions developed on the example of Estonian Wordnet

Abstract:

The purpose of this project was to automatically generate definitions for the terms in Estonian Wordnet that do not have existing definitions. The history of computer lexicons, as well as descriptions of wordnet lexicons and the various methods by which they are created, are presented in the theoretical section of the thesis. Further details on the creation, content and features of Estonian Wordnet are outlined. Additionally, guidance on the formalisation of the definitions is provided.

As a result of the project, a program was developed that produces definitions for terms that lack definitions in Estonian Wordnet. This was done using four different methods. It was concluded that definitions could be generated, but no method provided 100% correct definitions. Hence, all the generated definitions had to be verified. A total of 11,075 definitions were generated for the 18,731 missing definitions. The highest number of definitions (5469 definitions or ~50% of definitions) was generated based on the unique synset member method. The methods of similarity and the unique synset member worked with the best accuracy - 91% and 84% of the definitions were defined as fitting, respectively.

Keywords:

computer dictionary, Estonian Wordnet, automatic generation of definitions

CERCS: P170 Computer science, numerical analysis, systems, control

Sisukord

Sissejuhatus

Masinõpe on üha laialdasemalt kasutusel nii arvuti- ja andmeteaduses kui ka mitmes muus valdkonnas, et kiiremini andmeid analüüsida ja nende põhjal informeeritumaid otsuseid teha. Masinõpe on tehisintellekti rakendus, mis annab võimaluse automaatselt läbi kogemuste õppida. Masinõppe mudelid kasutavad õppimiseks ja kogemuse saamiseks erinevaid inimeste poolt ette antud või programmide poolt kogutud andmeid. Selleks, et arvuti oskaks ette antud andmeid analüüsida ja töödelda, peab neid regressiooni, klassifitseerimise, keeletöötluse jm tehnikate või meetoditega töötlema.

Üks olulisemaid masinõppe valdkondi andmete mõistmiseks on keeletehnoloogia. Keeletehnoloogia on kasutusele võetud mitmetes erinevates igapäevaelu mõjutavates valdkondades. Näiteks Google Mail kasutab keeletehnoloogia meetodeid selleks, et klassifitseerida emaili sisu põhjal, kas email kuulub peamisesse, suhtlusvõrgustiku või pakkumiste postkasti. Otsingumootorid, nt Google Search, kasutatavad keeletehnoloogiat selleks, et kasutajale sarnase otsingu käitumise põhjal asjakohased tulemused esile tuua. See võimaldab igal inimesel ilma suurema vaevata vajaliku infoni jõuda.

Keeletehnoloogia erinevad meetodid aitavad arvutil analüüsida ja töödelda loomulikku keelt ning seetõttu on keeletehnoloogia valdkond osutunud väga kasulikuks erinevate tõlkemasinate, juturobotite jne loomisel. Loomuliku keele semantiliseks analüüsiks on kasulik kasutada leksikaalset andmebaasi, mis sisaldab endas erinevate sõnade tähendusi ning nende sõnade vahelisi semantilisi suhteid, näiteks *wordnet'i*. Kasutades *wordnet*'is eksisteerivaid semantilisi suhteid on arvutil võimalik tuvastada näiteks, et „kass" on „koduloom" ning „koduloom" on „loom". Lisaks on erinevate keelte *wordnet*'id omavahel seotud keeltevahelise indeksiga, mistõttu on võimalik leida mõistepõhiseid tõlkevasteid. Kokkuvõttes aitab semantiliste suhete tuvastamine arvutil kombineerida efektiivselt omavahel struktureeritud ja mittestruktureerituid andmeid, mille põhjal õppida.

Wordnet tüüpi sõnastikke on loodud rohkem kui 200 keelele, kaasa arvatud eesti keelele (vt Global WordNet[1]). Eesti Wordnetis on hetkel (mai 2022) üle 91 700 mõiste ning üle 306 000 suhte, mis hõlmab eesti keele põhisõnavara (Eesti Wordneti koduleheküg[2]). Eesti Wordneti koostamiseks on peamiselt kasutatud EKI sõnastikke, erialasõnastikke, Vikipeediat või on kasutatud tõlkeid Princetoni WordNetist. Sõnad on *wordnet*-tüüpi

[1] Global WordNet Association koduleheküg. (i.a). http://globalwordnet.org/resources/wordnets-in-the-world/
[2] Eesti Wordneti koduleheküg. (i.a). https://www.cl.ut.ee/ressursid/teksaurus/?lang=et

sõnastikes jaotatud tähenduse alusel süno-hulkadesse. Selleks, et *wordnet*'e saaks võimalikult kasulikult rakendada, peab seal olema palju infot erinevate sünohulkade ja nende tähenduste kohta. Üheks Eesti Wordneti probleemiks on see, et paljudel (pool)automaatselt moodustatud sünohulkadel puuduvad seletused. Selle bakalaureusetöö eesmärk ongi automaatselt genereerida seletusi Eesti Wordneti sünohulkadele, millel seletus puudub, ja anda hinnang automaatselt koostatud seletuste korrektsusele.

Bakalaureusetöö esimeses peatükis antakse lühike ülevaade arvutileksikonide väljakujunemistest, olulisematest arvutileksikonide projektidest. Teises peatükis antakse ülevaade *wordnet*'i ajaloost, erinevatest koostamise viisidest. Lisaks antakse ülevaade Eesti Wordnetist, selle hetkeseisust ja probleemidest. Kolmandas peatükis kirjeldatakse, kuidas koostada ja formaliseerida sõnaseletust. Töö neljandas peatükis kirjeldatakse eesmärgi täitmiseks loodud rakendust ja rakenduse tööprotsessi. Viiendas peatükis analüüsitakse eesmärgi täitmiseks loodud rakenduse sooritust.

1. Arvutileksikonide ajalugu

Sõnastike koostamine arvuti abiga sai alguse 1960. aastatel, kui alustati suuremate teksti koguste ja korpuste digitaliseerimist. Korpuste põhjal on võimalik koostada sõnaloendeid ja erinevaid indekseid, mida kasutati leksikoloogilises analüüsis. Esimene tuntud arvutiga koostatud sõnastik on *Longman Dictionary of Contemporary English*, mis koostati 1978. aastal (Muischnek jt, 2003). Kaasaskantavad elektroonilised leksikonid tulid müüki samal aastal, nt Lexiconi poolt toodetud elektrooniline tõlkesõnastik LK-3000. LK-3000 sisaldas endas mitmeid tõlkesõnastikke, mille lähtekeeleks oli inglise keel ning tulemuseks oli võimalik saada vaste mitmes teises keeles, näiteks prantsuse, saksa või kreeka keeles (Nesi, 2008).

1980. aastatel pakuti esmakordselt välja idee hakata välja töötama leksikonide loomise ja leksikaalse materjali esitamise põhimõtteid. Sel ajal loodi iga rakenduse jaoks oma leksikon, mis tõi endaga kaasa dubleerimist ja asjata kulutusi. Üks põhimõtete väljatöötamise eesmärk oli selle vältimine (Muischnek jt, 2003).

Järgmine suur samm arvutileksikonide ajaloos toimus 1987. aastal, kui avaldati COBUILDi sõnastiku projekti raames koostatud sõnaraamat. Korpus, millel leksikon põhines, sisaldas endas rohkem andmeid inglise keele kohta kui kõik eelnevad korpused. Seetõttu oli COBUILDi sõnaraamat üks esimesi sõnaraamatuid, mis põhines igapäeva inglise kõne- ja kirjakeele näidetel (Collins Dictionary Language Blog[3]).

1977. aastal asutati Eestis arvutuslingvistika sektor ning seetõttu jõudis ka esimene arvuti Keele ja Kirjanduse Instituuti (praegune Eesti Keele Instituut ehk EKI). Esimene arvutiga loodud eestikeelne sõnastik „Väike murdesõnastik I“ ilmus 1982. aastal. Sellega algas Eestis sõnaraamatute sisestamine arvutisse, eesmärgiga moodustada ühine sõnastike süsteem. (Tender ja Viikberg, 2017).

1980-ndatel leiti, et oleks kasulik võtta kasutusele andmebaasi vorm. Leiti, et andmebaasi kasutamisega oleks võimalik materjali automaatselt klassifitseerida ja sõnaseletusi genereerida. Seetõttu hakati arvutis olevaid sõnastikke kasutama semantiliste hierarhiate ehitamiseks. Nii tekkisid leksikaalsed andmebaasid, millel põhineb *wordnet* (Muischnek jt, 2003). Edasi tutvustatakse *wordnet*-tüüpi sõnastiku ülesehitust.

[3] The history of COBUILD. Collins Dictionary Blog. https://blog.collinsdictionary.com/the-history-of-co-build/

2. WordNet

WordNet on leksikosemantiline andmebaas, kus sõnad on sünonüümide põhjal grupeeritud sünohulkadesse ning erinevad sünohulgad on omavahel seotud erinevate semantiliste suhetega. Näiteks sünohulk „tamm, tammepuu" on seotud sünohulgaga „puu" hüponüümia ja hüperonüümia suhte kaudu (tamm on puu, aga puu ei ole ainult tamm). Esimese *wordnet*'i arendamist alustati 1980. aastate keskpaigal inglise keeles Princeton'i Ülikoolis (Princeton WordNet[4]). Tänapäeval on *wordnet*'e loodud rohkem kui 200 keeles (vt Global WordNet).

2.1 WordNet'ist üldiselt

Wordnet'i algne eesmärk oli luua mudel kirjeldamaks, kuidas sõnad inimese peas on omavahel seotud (Fellbaum, 1998). Algselt loodi *wordnet* psühholoogide ja keeleteadlaste uurimistulemuste kirjeldamiseks, kuid tänapäeval kasutavad *wordnet*'i pigem keeleteadlased (Kilgarriff, 2000). Lisaks on tänapäeval *wordnet* esile tõusnud ka infotehnoloogia valdkonnas, et arvuti saaks keeleandmete põhjal teha semantilisi järeldusi (Orav jt, 2011). Näiteks on võimalik kasutada osa-terviku suhteid (köök on kodu osa) juturobotites ja muudes rakendustes, hüponüümia ja hüperonüümia suhet (hüponüümide „koduloom" ja „metsloom" hüperonüümiks on loom) saab kasutada laiemaks või kitsamaks infootsinguks. Lisaks kasutatakse *wordnet*'ist saadavat infot ontoloogiate koostamiseks – seega kõik rakendused, kus osutuvad vajalikuks hierarhiad, saaksid kasutada *wordnet*'i. Teiseks on erinevate keelte *wordnet*'id omavahel seotud keeltevahelise indeksiga, mistõttu on võimalik leida mõistepõhiseid tõlkevasteid ja liikuda võrgustikus vähese vaevaga ühest keelest teise (Orav jt, 2015).

Paljude erinevate *wordnet*'ide aluseks on võetud Princetoni WordNeti põhimõtted, aga nende koostamise strateegiad on olnud erinevad. Suures pildis jaotuvad strateegiad kolme kategooriasse: käsitsi, poolautomaatselt ja automaatselt (Orav jt, 2015). Käsitsi, mis on ühtlasi üks populaarsemaid viise, tähendab seda, et *wordnet* koostatakse vastava keelte üks-keelsete sõnaraamatute põhjal ning seejärel ühendatakse see Princetoni WordNetiga. Sel viisil on oma *wordnet*'i koostanud näiteks taanlased (DanNet). Poolautomaatse strateegia korral on kasutatud nii vastava keelte sõnaraamatutest saadud informatsiooni, kui ka tõlkeid Princetoni WordNetist (või mingist muust *wordnet*'ist), mis on oma keelele vastavaks kohandatud. Poolautomaatselt on koostatud näiteks Eesti Wordnet. Automaatselt *wordnet*'i

[4] Princetoni Wordneti koduleheküg. (i.a). https://wordnet.princeton.edu/

koostamise korral kasutatakse mõnda teist *wordnet*'i (peamiselt Princetoni WordNeti) ning tõlgitakse see vastavalt oma keelele ümber. Niimoodi on koostanud oma *wordnet*'i soomlased (FinnWordNet) (Pedersen jt, 2013).

2.2 Eesti Wordnet

Orav jt (2015) on oma artiklis välja toonud, et Eesti Wordnet (EstWN) sai alguse 1995. aastal EuroWordNeti projekti raames (Orav jt, 2015). Selle alustamisel järgiti nii Princetoni WordNeti kui ka EuroWordNeti põhimõtteid. Eesti Wordneti koostamist alustati EuroWordNeti projekti raames saadud baasmõistete tõlkimisega, mida laiendati korpuste sagedusloendite järgi. Kõige suurem Eesti Wordneti areng on toimunud Eesti riikliku keele-tehnoloogia programmi raames, mis sai alguse 2007. aastal ning kestab tänapäevani (Orav jt, 2015).

Seda väidet toetab ka Kahuski ja Videri (2017) uurimistöö, mille raames käsitleti Eesti Wordneti suuremaid iga-aastaseid uuendusi vahemikus 1998 kuni 2016. Töös toodi välja, et Eesti tesauruses on perioodil 2007–2016 toimunud keskmiselt suurem sünohulkade arvu tõus võrreldes 1998–2007 perioodiga. Kõige suurem muutus toimus 2010. aastal, kui lisati võrreldes 2009. aastaga umbes 15 500 sõna. Allika järgi on Eesti Wordneti sünohulkade arv mitte lineaarselt kasvanud, vaid pigem S-kujuliselt, kus 2000ndate aastate alguses oli kasv rahulik ning järsk tõus toimuski 2010. aastal. Peale seda pole nii hüppelist sünohulkade arvu muutust tulnud (Kahusk ja Vider, 2017).

Eesti Wordnet koosneb erinevatest sõnaliikidest (adjektiivid, substantiivid, verbid ja adverbid) ning kõik sõnad on koondatud sünohulkadesse täis- ja lähisünonüümide alusel (Orav jt, 2015). Eesti Wordneti ametliku koduleheküljе järgi on 2021. aasta oktoobrikuu andmete põhjal seal ligi 91 700 mõistet ehk sünohulka, 148 000 sõna ja 306 000 semantilist suhet. Lisaks on osad mõisted seotud keeltevahelise indeksi põhjal Princetoni Wordnetiga (Eesti Wordneti koduleheküлg).

Eesti Wordnetis keskendutakse võrdselt nii sõnade arvu suurendamisele kui ka nende kvaliteedile. Hea kvaliteedi tagamiseks peavad uued sõnad vastama kolmele tingimusele:

- peavad piisavalt sagedaselt esinema tavapärastest tekstides;
- peavad olema olulised erinevatele rakendustele;
- erinevate *wordnet*'ide ühendamine peab olema võimalik, selleks et oleks võimalik võrrelda teatud keeles mingi mõiste seoseid teiste mõistetega (Orav jt, 2015).

Sõnade arvu on suurendatud enamasti käsitsi, aga on proovitud automaatselt (Orav jt, 2015). Sõnade automaatseks lisamiseks Eesti Wordneti peab kokku panema sõnastikke, mis eristavad/seletavad ükskeelseid tähendusi ning sünonüüme. Automaatselt sõnade lisamine õnnestus *mine*-tegevusnimede ning *ja*-tegijanimede puhul. Selleks kasutati *ma*-tegevusnimest tuletamise meetodit (Orav jt, 2011). Automaatselt sõnade lisamisega kaasnes ka mitmeid probleeme. Näiteks paljude sünohulkade puhul oli kahe tähenduse erinevus ainult kontekstist sõltuv. Esines ka olukordi, kus seletus viitas kahele erinevale tähendusele. Lisaks saadi ka sünohulki, mida mõjutab sooline eripära (Villem 2009). Näiteks nagu „veis", „pull", „lehm", kus „veis" peaks olema „pulli" ja „lehma" ülemmõiste. Nendest probleemidest järeldati, et Eesti Wordneti on mõttekam täiendada käsitsi, sest automaatselt lisatud sünohulkade käsitsi kontrollimisele kulus rohkem aega (Villem 2009).

2.3 Eesti Wordneti probleemid

Üheks suuremaks probleemiks nii Eesti Wordnetis kui ka üldiselt *wordnet*'is on esmatähtsate semantiliste suhete määramine ning nende ühene mõistmine. Kokku on Eesti Wordnetis 51 erinevat suhtetüüpi (Orav jt, 2015). Peamised suhted, mille erinevalt mõistmine Eesti Wordnetis on põhjustanud probleeme, on: sünonüümia, hüponüümia ja hüperonüümia, holonüümia ja meronüümia, rollisuhe, hägussuhe. Suuremates rahvusvahelistes *wordnet*'i projektides on probleemi lahendamiseks ette antud kindlad suhted, mida *wordnet* peab endas sisaldama. Sel viisi proovitakse vähendada suhete olemusest tulenevat arusaamatust. Näiteks Eesti Wordnet on saanud oma semantiliste suhete nimekirja kaasa EuroWordneti projektist (Orav jt, 2015).

Paljud *mine*-liitelised nimisõnad ning *ja*-tegijanimed on EstWNi lisatud automaatselt ja neil seetõttu puuduvad seletused. *Mine*-liiteliste nimisõnade puhul saab kasutada *ma*-tegevusnime seletust, sest *mine*-liide ei muuda definitsiooni sisu. Sel viisil *mine*-liiteliste nimisõnade defineerimise korral on võimalik üle kanda ka olemasolevad semantilised suhted (Orav jt, 2015). Näiteks sõna „algatamine" puhul saaks kasutada sõna „algatama" seletusest tuletatud seletust „millegi algatamine, arutlusele võtmine, nt probleemi püstitamine". *Ja*-tegijanimede puhul saab kasutada sama lähenemist. Sel juhul tuleb kirjutada seletust stiilis „keegi, kes midagi teeb" või „miski, mis midagi teeb". Näiteks sõna „algataja" puhul tuleb kirjutada seletus „keegi, kes midagi algatab, arutlusele võtab, nt probleemi püstitab".

Lisaks erinevatele suhete tõlgendamise probleemidele on Eesti Wordnetis raskusi põhjustanud ka liitsõnad. Eesti keeles on väga palju liitsõnu ja püsiühendeid ning seetõttu

peavad Eesti Wordneti koostajad otsustama, milliseid liitsõnu Eesti Wordneti lisada. Liitsõnade lisamisel tuleb jälgida, kui sageli need sõnad keeles esinevad (Orav jt, 2011). Näiteks liitsõna „raudtee" on eesti keeles piisavalt populaarne ning igapäevane sõna, et lisada Eesti Wordneti, aga näiteks „elevandilondikondiüdi" on pigem harva esinev sõna. Seetõttu pole seda sõna EstWNi mõtet lisada.

Sarnane probleem on ka püsiühenditega ehk idioomidega. Eesti Wordneti sünohulkadesse saab lisada ainult idioome, mille vorm ei muutu kunagi ja tähendus on kõigile üheselt mõistetav (Orav jt, 2011). Näiteks idioom „auku pähe rääkima" võib lisada sünohulka „veenma", sest see on kõigile üheselt mõistetav. Kui lisada kõik idioomid Eesti Wordnetti võib tekkida nii süntaksi kui ka semantilisi vigu – nt ei või lisada eitavas vormis püsiühendeid, sest need kaotavad jaatavas kõneviisis oma tähenduse (Orav jt, 2011).

3. Sõnaseletuste formaliseerimine

Sünohulga sisu avavad Eesti Wordnetis nii semantilised suhted kui ka sõnaseletus. Seletuste kirjutamisel võiks võimalusel jälgida kahte olulist eesmärki: see peab andma edasi selle vastava sünohulga tähenduse ning konteksti, milles võib sünohulga liikmeid kasutada (Uiboaed, 2005). Eesti Wordnetis on seletuste koostamisel kasutatud peamiselt EKI sõnastikke, erialasõnastikke, Vikipeediat, aga osad sünohulkade seletused on koostajad ise kirjutanud.

Sõnaseletust kirjutades on võimalik kasutada kahte elementi – soomõistet ja liigierisust:

- lähim soomõiste kui ülemmõiste määrab, millises semantilises kategoorias seletatav sõna on;
- vähemalt ühest lisa eristustunnusest, mis eristab seda sõna teistest selle semantilise kategooriate liikmetest (Meos, 2003).

Näiteks nii saab defineerida sõna „kass" lausega „kaslaste hulka kuuluv koduloom", kus soomõiste on „kaslane" ning liigierisus on „koduloom". Selline meetod töötab hästi enamuse nimisõnade korral, mis viitavad mingile kindlale objektile (näiteks kast, lusikas) või ka tegusõnade puhul, mis on seotud loomise või tegemisega (Atkins ja Rundell, 2008). Näiteks „välja ehitama" ja „ümber ehitama" saab defineerida ehitamise kaudu, mis on omakorda hüponüüm loomisele ja tegemisele.

Määrsõnade ja omadussõnade puhul ei saa soomõiste ja liigierisuse kaudu seletust kirjutada, sest nende sõnade puhul on sageli võimatu leida ülemmõistet. Seetõttu mõtlevad sõnaseletuste autorid määrsõnade ja omadussõnade sõnaseletused ise välja, järgimata kindlat meetodit (Atkins ja Rundell, 2008). Näiteks sõna „ilus" on Eesti Keele Instituudi ühendsõnastikus defineeritud „välimuselt meeldiv, vaadates rahuldust pakkuv, väga kena" (EKI ühendsõnastik 2022). Antud seletuses puudub selge ülemmõiste kui ka liigierisus.

Sõnaseletuse sisu sõltub peamiselt sellest, kellele on sõnastik suunatud (Atkins ja Rundell, 2008). Näiteks sõna „absorptsioon" puhul piisab juhuslikule lugejale, seletus „valguse neeldumine", kuid eriala huviline eeldab definitsioonis põhjalikumat kirjeldust. Seetõttu tuleb arvestada, millises tekstis võib sõnastiku kasutaja mingi sõnaga kokku puutuda ja kui põhjalik seletus tuleks sõnale anda. Arusaadava ning piisavalt põhjaliku seletuse saavutamiseks järgitakse järgmiseid soovitusi:

- sõnaseletused peavad olema piisavalt lühikesed ja selgesti arusaadavad, selleks et lugeja ei peaks kasutama teist sõnastikku sõnaseletuse sees olevate sõnade tähenduste uurimiseks (Atkins ja Rundell, 2008);
- sünohulka ei tohi defineerida sünohulga liikme kaudu;
- seletus ei tohi olla ühesõnaline;
- seletus peab sisaldama piisavalt informatsiooni, et lugeja saaks aru, mis kontekstis seda kasutada (Atkins ja Rundell, 2008);
- sõnaseletus peab minimaalselt sisaldama nii palju informatsiooni, et lugeja saaks aru, mida antud sõna tähendab selles kontekstis, kus ta sellega kokku puutus (Atkins ja Rundell, 2008).

Ülaltoodud soovitusi kasutatakse nii sõnade puhul, mille seletust saab kirja panna kasutades soomõistet ja liigierisust, kui ka sõnade puhul, millel puudub ülemmõiste. *Wordnet*-tüüpi sõnastikus aitavad semantilised suhted ka kindlasti tähendust avada. Kuna siinse töö eesmärk on Eesti Wordneti sünohulkadele genereerida seletusi ning neid kontrollida, siis järgnevalt tutvustatakse meetodeid, kuidas seletusi genereeriti.

4. Meetodi kirjeldus

Bakalaureusetöö raames koostatud rakenduse eesmärgiks oli genereerida puuduvad seletused Eesti Wordneti sünohulkadele. Sõnaseletuste saamiseks kasutati mitmeid allikaid:

- Eesti Keele Instituudi sõnastiku- ja terminibaasisüsteemi (Ekilex[5]).
- Kuna Eesti Wordnet on ära seotud Princetoni Wordnetiga, siis sai mõnede sünohulkade puhul juurde lisada Princetoni Wordneti ingliskeelsest seletusest tõlgitud seletuse. Seletuse tõlkimiseks kasutati Tartu NLP Neurotõlke[6] rakendust.

Sõnaseletusi genereeriti neljal erineval viisil: sünonüümide sarnasuse, unikaalse sünohulga liikme, *mine*-tegevusnime algvormi ning *ja*-tegijanime algvormi põhjal. *Mine*-tegevusnimede ning *ja*-tegijanimede sünohulgad on Eesti Wordneti lisatud automaatselt ilma seletusteta, mistõttu keskenduti nende sünohulkade seletuste genereerimisele eraldi. Lisaks on need liited eesti keeles laialdaselt kasutusel, aga paljudes sõnastikes puuduvad nende sõnade seletused. Seetõttu pidi sõnaseletuse saamiseks tegema Ekilexi päringu sõna algvormiga (*ma*-tegevusnimega). Algvormi seletuse õigesse vormi teisendamiseks kasutati EstNLTK[7] funktsioone. Lisaks pidi iga genereeritud seletus vastama kolmele tingimusele: seletus ei või olla ühesõnaline, sünohulk ei tohtinud olla defineeritud sünohulga liikme kaudu ning seletus ei võinud olla juba mõne teise sünohulga liikme seletuseks (vt ptk 3). Järgnevalt on antud ülevaade sõnaseletuse genereerimise protsessist.

4.1 Rakenduse kirjeldus

Rakendus kasutas sõnaseletuse genereerimiseks põhiliselt nelja erinevat meetodit. Meetodid on välja toodud läbimise järjekorras, järgmist meetodit kasutati ainult juhul, kui eelmine meetod tulemust ei andnud. Igas meetodis kontrolliti viimase sammuna, kas Princetoni Wordnetis leidub võrdne (eg synonym suhetega) sünohulk. Juhul kui leidus, kasutati Princetoni Wordnetist saadud sünohulga seletuse tõlkimiseks TartuNLP Neurotõlget ning lisati lõpptulemusele juurde, et hiljem võrrelda kumb seletus sobib paremini.

I samm – sünohulga seletuse leidmine sarnasuse põhjal

Esiteks määratakse EstWNi sünohulk, millel seletus puudub ning tehakse iga sünohulga liikmega eraldi päring Ekilexi. Ekilexist tagastatakse iga sõna vastav(ad) unikaalne(sed)

[5] EKI sõnastiku- ja terminibaasisüsteem (Ekilex). https://ekilex.ee/
[6] TartuNLP Neurotõlge. (i.a). https://translate.ut.ee/
[7] EstNLTK dokumentatsioon. (i.a). https://estnltk.github.io/

ID-d ning sünonüümid. Unikaalne ID näitab Ekilexis, milline seletus sellele sõnale vastab (ühel sõnal võib olla mitu seletust). Iga sünonüümi hulga puhul arvutatakse välja selle sarnasus vastava EstWNi sünohulgaga. Valiti välja seletus, mis vastas eelnevalt kirjeldatud tingimustele (vt ptk 4) ning mille sünohulk oli Ekilexist saadud sünonüümide hulgaga kõige sarnasem.

Näide:

1. Valitud puuduva seletusega EstWNi sünohulk: [humaansus, humanism, inimlikkus, inimsõbralikkus]
2. Ekilexist saadud sünonüümide põhjal arvutatud sarnasused: {27386: 0.2, 600945: 0.25, 27396: 0.0, 27397: 0.0, 27398: 0.0, 339714: 0.0, 451934: 0.0, 591145: 0.0, 600946: 0.0}
3. Sarnaseim sünonüüm: [600945]
4. Saadud seletus: inimsus, inimsõbralikkus, inimväärikuse austamine
5. Neurotõlge (eng → est): kaastunde või teistega arvestamise kvaliteet (inimesed või loomad)

Selle meetodi negatiivseks küljeks on, et Ekilexis ei ole paljude sõnade juures sünonüüme välja toodud, mille tõttu pole seda meetodit sageli võimalik kasutada.

II samm – sünohulga seletuse leidmine unikaalse sünohulga liikme põhjal

Kui eelnev viis tulemust ei andnud, siis eemaldati unikaalse sünohulga liikmete loetelu saamiseks kõik sünohulga liikmed, mis esinesid lisaks ka mõnes teises sünohulgas. Kasutades unikaalsetest sünohulga liikmetest koosnevat sünohulka oli kõige suurem tõenäosus, et sünohulgale leitakse õiges kontekstis seletus. Järgmisena tehti alles jäänud sõnadega ükshaaval päring Ekilexi. Peale igat päringut kontrolliti, kas Ekilexist saadud seletus vastas tingimustele. Kui seletus vastas tingimustele, tagastati vastav seletus. Juhul kui seletust ei leitud, kasutati sünohulga järgmist liiget, kuni leiti sobiv seletus.

Näide :

1. EstWNist saadud sünohulk: [aadamaülikonnas, alasti, eevaülikonnas, ihualasti, paljas, porgandpaljas]
2. Unikaalne sünohulk: [aadamaülikonnas, eevaülikonnas, ihualasti, porgandpaljas]
3. Sõna, millega tehti päring: ihualasti
4. Saadud seletus: täiesti alasti, ühegi riidehilbuta
5. Neurotõlge (eng → est): täiesti riieteta

Antud meetodi negatiivseks küljeks on see, et ei saa genereerida seletusi sünohulkadele, mille kõik liikmed esinevad ka teistes sünohulkades. Näiteks sünohulgale „ahistamine, kimbutamine" ei saa sel viisil seletust leida, sest „ahistamine" esineb veel sünohulkades „ahendamine, ahendus, ahenemine, ahistamine", „ahistamine", „ahistamine, ängistamine, koormamine, painamine, piinamine, rõhumine, rusumine, vaevamine" ning sõna „kimbutamine" esineb veel sünohulgas „kimbutamine, kimbutus, kiusamine, kiuste, piinamine, vaevamine". Seega jääb puudu tähenduste eristamisest.

III samm – *mine*-tegevusnimedest koosneva sünohulga seletuse leidmine

Juhul kui sünohulk koosnes *mine*-tegevusnimedest ning eelnevad meetodid tulemust ei andnud, kasutati sõnaseletuse saamiseks *mine*-tegevusnime meetodit. Alguses kontrolliti, kas vähemalt 50% on ühe sünohulga liikmetest *mine*-liitega. Seejärel jäeti alles ainult unikaalsed sünohulga liikmed. Järgmiseks asendati sünohulga esimese liikme *mine*-liide (nt jooksmine) *ma*-liitega (nt jooksma) ning tehti sellega päring Ekilexi. Seletuse leidumisel teisendati seletus õigesse vormi kasutades EstNLTK funktsionaalsust. Ainsuse osastavas ja lühikeses sisseütlevas käändes sõnad teisendati ainsuse omastavasse käändesse ning mitmuse osastav teisendati mitmuse omastavasse käändesse. Juhul kui sünohulga esimese liikmega seletust ei leitud, kasutati sünohulga järgmist liiget, kuni leiti sobiv seletust.

Seletuse korrigeerimiseks kasutati ka *mine*-tegevusnimede seletuste genereerimise meetodil TartuNLP neurotõlget eesti keelest eesti keelde, sest kõikide Ekilexist saadud seletuste puhul ei saadud seletust automaatselt korrektseks muuta. Neurotõlke kasutamine võimaldas saada parema tulemuse ja keelevead parandada.

Näide:

1. EstWNist saadud sünohulk: [absolutiseerimine, absoluudistamine, absoluutimine]
2. Sõna, millega tehti päring Ekilexi: absolutiseerima (sünohulga esimese liikme algvorm)
3. Ekilexist saadud seletus: täiuslikuks, kõikehõlmavaks vms taoliseks pidama või muutma
4. Muudetud seletus: täiuslikuks, kõikehõlmavaks vms taoliseks pidamine või muutmine
5. Neurotõlge (est → est): ideaalseks, kõikehõlmavaks või muuks taoliseks pidamine või muutmine
6. Neurotõlge (eng → est): puudub

Puuduseks on see, et sel viisil sõnaseletuste genereerimine võib anda eesti keele iseärasuste ja keerukuse tõttu keeleliselt ebakorrektse seletuse.

IV meetod – *ja*-tegijanimedest koosneva sünohulga seletuse leidmine

Juhul kui sünohulk koosnes *ja*-tegijanimedest ning Ekilexist seletust ei saadud, kasutati sõnaseletuse saamiseks *ja*-tegijanime meetodit. Alguses kontrolliti, kas vähemalt 50% sünohulga liikmetest on *ja*-liitega. Seejärel jäeti alles ainult unikaalsed sünohulga liikmed. Järgmiseks asendati sünohulga esimese liikme *ja*-liide (nt jooksja) *ma*-liitega (nt jooksma) ning tehti sellega päring Ekilexi. Seletuse leidumisel teisendati seletus õigesse vormi kasutades EstNLTK funktsionaalsust. Algvormis tegusõnad teisendati oleviku 3. isiku ainsuse aktiivi. Samuti lisatakse teisendatud seletuse algusesse „keegi, kes", sest suure tõenäosusega on tegija elusolend. Juhul kui sünohulga esimese liikmega seletust ei leitud, kasutati sünohulga järgmist liiget, kuni leiti sobiv seletus.

Näide:

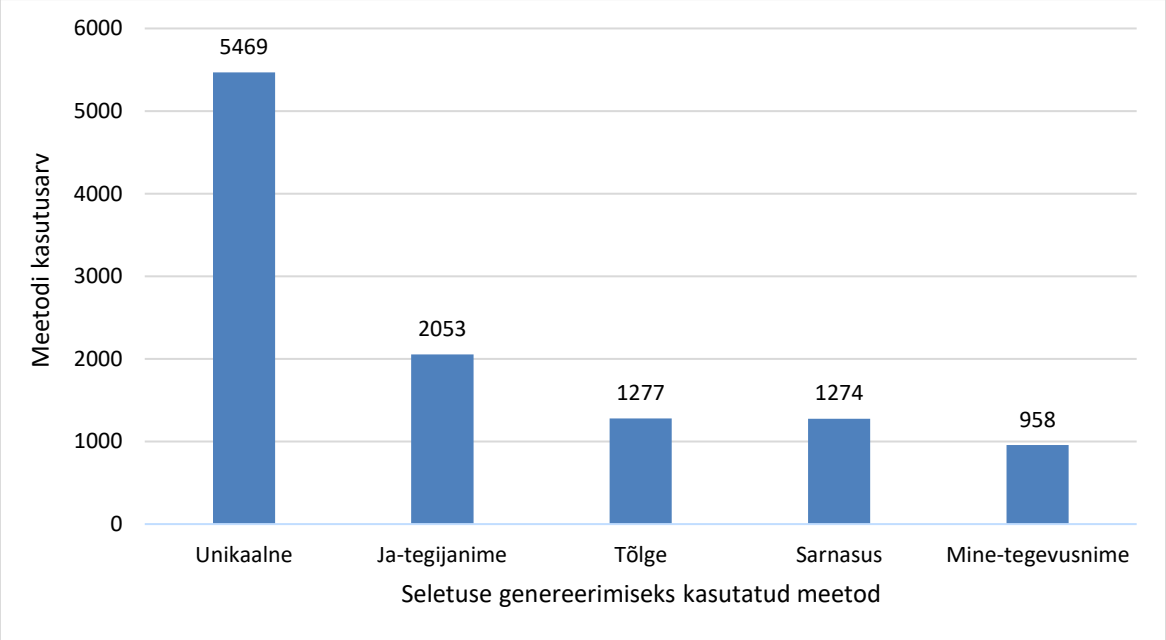
1. EstWNist saadud sünohulk: [aktsepteerija, leppija, nõustuja, soostuja]
2. Sõna, millega tehti päring Ekilexi: soostuma (sünohulga viimase liikme algvorm)
3. Ekilexist saadud seletus: milleski järele andma, millegagi (lõpuks) nõustuma
4. Muudetud seletus: keegi, kes milleski järele annab, millegagi (lõpuks) nõustub
5. Neurotõlge (eng → est): isik, kes on nõus

Sel viisil sõnaseletuste genereerimine võib anda eesti keele iseärasuste ja keerukuse tõttu keeleliselt ebakorrektse seletuse.

Kokkuvõttes pole neid meetodeid kasutades sõnaseletuse genereerimine perfektne ning esineb nii positiivseid kui negatiivseid külgi. Näiteks on positiivne see, et antud meetoditega sõnaseletuse genereerimine tagab sõnaseletuste mittekordamise. Lisaks on sünonüümide sarnasuse ja unikaalse sünohulga liikme põhjal saadud seletused keeleliselt korrektsed. Negatiivne on see, et *mine*-tegevusnime või *ja*-tegijanime meetoditega genereeritud seletused võivad olla keeleliselt ebakorrektsed. Kuna eesti keeles on palju kahetähendus-likke sõnu, siis ei olnud algvormis oleva seletuse õigesse vormi panemine alati võimalik. Loodud rakenduse kõige suuremaks veaks on ülem- ja alamsuhete mittearvestamine, mis-tõttu võivad osad sünohulga seletused tulla liiga spetsiifilised. Töö raames koostatud rakendus ning genereeritud seletused on kättesaadaval https://github.com/markovkristo/Se-letuste-automaatne-genereerimine-ja-kontrollimine-Eesti-Wordneti-naitel.

5. Genereeritud seletuste analüüs

Eesti Wordnetis puudus kokku 18 731 sünohulgal seletus, millest 2532 sünohulka olid *mine*-liitega ning 3566 sünohulka olid *ja*-liitega. Kokku genereeriti 11 075 sõnaseletust ning kõige tulemuslikumaks meetodiks osutus sünohulga seletuse leidmine unikaalse sünohulga liikme põhjal (joonis 1). Lisaks genereeriti eraldi meetodiga seletused 48-le numbritest koosnevatele sünohulkadele, nt sünohulgale „17, seitseteist, seitseteistkümmend“ või „17., seitsmeteistkümnes, XVII“. Numbritele genereerituid seletusi analüüsis ei arvestatud, sest tulem oli liiga väike. 7656 sünohulgale nende meetoditega tulemust ei saadud. 453 korral sai *mine*-liitelistele sünohulkadele seletust genereerides teha sünohulga liikmega päringu Ekilexi (tabel 1). 958 korral pidi *mine*-liite teisendama *ma*-liiteks ehk kasutama *mine*-tegevusnime meetodit. 351 korral sai *ja*-liitega sünohulkadele seletuse tehes teha vastava sünohulga liikmega päringu Ekilexi, 2053 korral pidi *ja*-liite teisendama *ma*-liiteks ehk kasutama *ja*-tegijanime meetodit (tabel 1).



Joonis 1. Erinevate meetodite kasutusarv seletuste genereerimiseks

Tabel 1. *Mine*- ning *ja*-liitega sünohulkade puhul kasutatud meetodite arv

	Sarnasus	Unikaalne	*Mine*-tegevusnimi	*Ja*-tegijanimi	Tõlge
Mine-liitega	74	379	958	0	125
Ja-liitega	61	290	0	2053	85

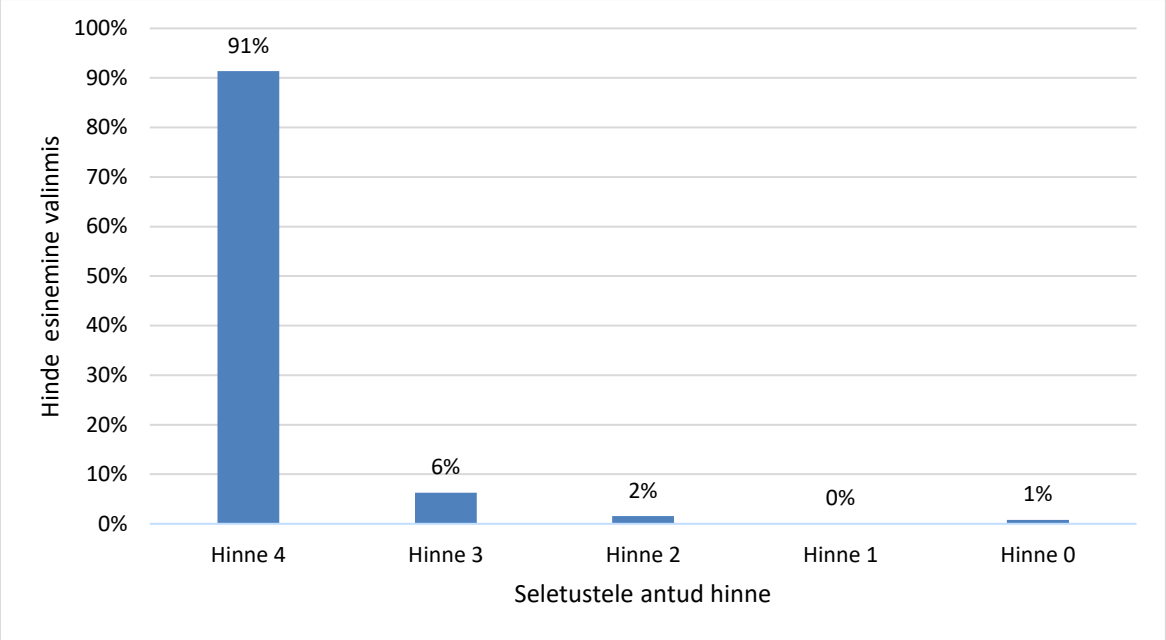
Kuna iga meetodi kasutuste arv oli erinev, siis valiti juhuslikult 20% iga meetodi poolt genereeritud seletustest ning hinnati nende seletuste sobivust. Hinnati vaid neid seletusi, kus saadi seletus ainult Ekilexist või tõlkides Princetoni Wordnetist saadud seletust. Seletuse sobivust hinnati skaalal nullist neljani, kus neli tähistas sünohulgale sobivat seletust ning null tähistas, et genereeritud seletus ei sobi vastavale sünohulgale. Genereeritud seletuste analüüs on esitatud meetodite töötamise järjekorras.

5.1 Sarnasuse meetodiga genereeritud seletuste analüüs

Sarnasuse meetodit kasutati 1274 korral ning analüüsitava valimi suurus oli 255 sõnaseletust (joonis 1). Saadud seletuste keskmiseks hindeks oli 3,81. Sobiv sõnaseletus (hinne 4) genereeriti valimis antud meetodiga 91% kordadest (joonis 2). Näiteks genereeriti sünohulgale „euroliit, Euroopa Liit" seletus „Euroopas asuvate riikide majanduslik ja poliitiline ühendus", milles ei esine ühtegi kirjaviga, on sobiva pikkusega ning seletab sünohulga tähenduse piisavalt lahti.

Hinded 3 ja 2 anti enamasti seletustele, mis olid kas liiga pikad või lühikesed või ei kirjeldanud kogu sünohulka. Näiteks sünohulgale „pereheitmine, sülemlemine" genereeritud seletusele „mesilaspere loomulik paljunemisviis, mis toimub suve esimesel poolel. Enne uue mesilasema koorumist lendab vana mesilasema koos osa pere liikmetega tarust välja. Nad ringlevad õhus ja laskuvad kuhugi" anti hinne 3, sest seletus on liiga pikk. Antud sünohulgale oleks piisanud seletusest „mesilaspere loomulik paljunemisviis, mis toimub suve esimesel poolel". Hinne 2 anti näiteks seletusele „väärismetallist esemete valmistamisel üks vanimaid ja enamkasutatavaid töövõtteid, mis põhineb hõbeda tehnilistel omadustel, s.o. taotavusel, elastusel ja venitatavusel (Tamla 1998 24)", mis genereeriti sünohulgale „sepistamine, sepistus, sepitsemine, tagumine, taondamine, väljatagumine". Genereeritud seletus on antud sünohulga jaoks liiga eriala spetsiifiline.

Ainult 1% valimi genereeritud sõnaseletustest ei sobinud (joonis 2). Näiteks genereeriti sünohulgale „hävimine, hukkumine" seletus „(asja või vara kohta)". Saadud seletus genereeriti seetõttu, et Ekilexis oli sõna „hävimine" all mitu selgitust ning ainult selle sõnaseletuse juurde oli välja toodud „hukkumine", kui hävimise sünonüüm, mistõttu valiti seletus „(asja või vara kohta)".



Joonis 2. Sarnasuse meetodiga genereeritud sõnaseletuste hinnete jaotus

Kokkuvõttes on sarnasuse meetodi eksimisprotsent väike, aga sellegi eest tuleks seletused enne EstWNi lisamist üle kontrollida, sest seletused võivad olla liiga lühikesed, pikad või ebatäpsed. Kõige paremini saab sarnasuse meetodiga genereerida sobivaid seletusi sünohulkadele, millel on kaks liiget – 57% hinde 4 saanud seletustest olid kaheliikmelised. 24% sobivatest seletustest genereeriti sünohulkadele suurusega kolm. Sellest võib järeldada, et mida rohkem sünohulga liikmeid on, seda väiksem on tõenäosus genereerida sünohulgale sobiv seletus.

5.2 Unikaalse sünohulga liikme põhjal genereeritud seletuste analüüs

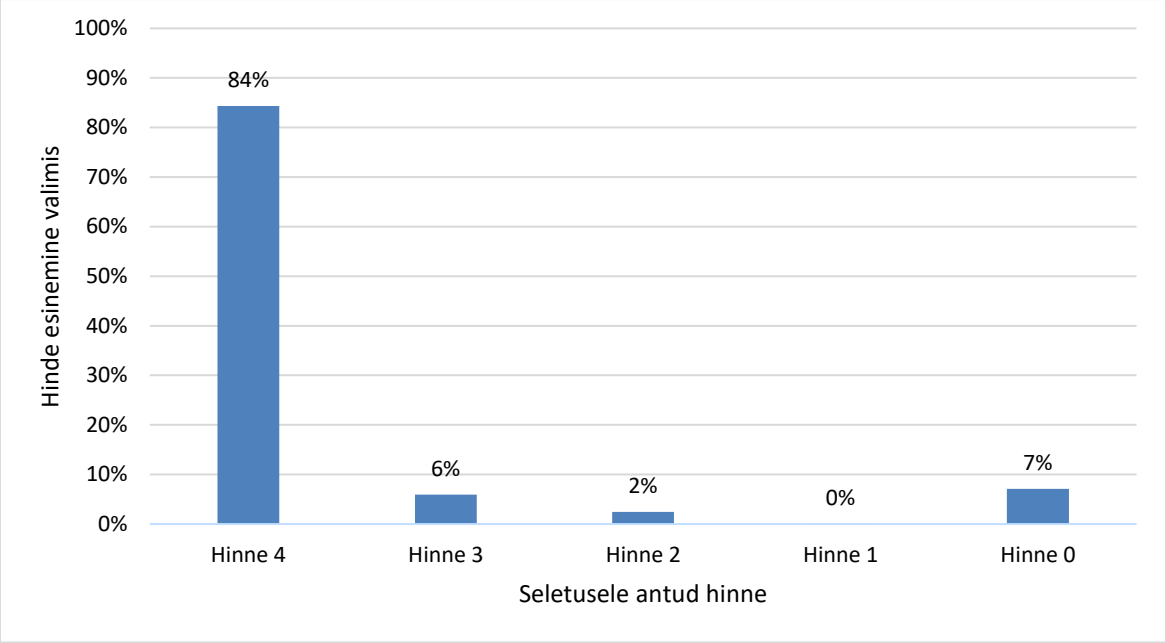
Unikaalse sünohulga liikme põhjal genereeriti seletus 5469 korral ning analüüsitava valimi suurus oli 1094 (joonis 1). Analüüsitavate seletuste keskmiseks hindeks oli 3,60. Hinde 4 saanud seletusi oli valimis kokku 84% (joonis 3). Näiteks genereeriti sünohulgale „polüvitamiin" seletus „mitut vitamiini sisaldav preparaat". Genereeritud seletuses ei esine ühtegi kirja- ega lausestusviga, see on sobiva pikkusega ning kirjeldab sünohulka piisavalt hästi.

Hinne 3 anti enamasti seletustele, mis olid kas liiga pikad, lühikesed või seletuses esines sünohulga liige. Näiteks sünohulgale „sügavdamine, süvendamine, süvendus" genereeriti seletus „maapinna, merepõhja vm pinnase sügavamaks uuristamine, millegi süvendamine", mis aga sisaldab endas sünohulga liiget. Seletus sobiks, kui eemaldada seletusest „millegi süvendamine". Sobiv seletus oleks sel juhul „maapinna, merepõhja vm pinnase

sügavamaks uuristamine". Lühikene seletus, mis ei kirjeldanud sünohulka piisavalt hästi oli näiteks „kõrgepingega elektriliin", mis genereeriti sünohulgale „kõrgepingeliin". Antud seletuses võiks lahti ka seletada, mida tähendab „kõrgepingega" ehk sobivam seletus oleks „elektriliin, mille elektripinge on kõrgem kui 1000 volti".

Hinde 2 said seletused, mis ei kirjeldanud sünohulka piisavalt hästi ning seetõttu nõuaks rohkem muutmist, et nendest saada sobiv seletus. Hinde 2 saanud seletusi oli valimis kokku 2%. Näiteks sünohulgale „murdeleksika, murdesõnavara" genereeriti seletus „murdekeeles esinevad sõnad". Saadud seletus ei kirjelda antud sünohulka õigesti, kuid sünohulgast ja seletusest võib välja lugeda, et seletus peaks olema „teatmeteos, mis annab infot murdekeeles esinevate sõnade kohta". Hinne 2 anti ka seletustele, milles esines nii sünohulga liige ja mis olid liiga lühikesed või pikad. Näiteks seletus „infotehnoloogia asjatundja, IT-spetsialist", mis genereeriti sünohulgale „infotehnoloog, IT-spetsialist". Selle sünohulga seletuses võiks lühidalt kirjeldada, mis valdkond infotehnoloogia on. Parem seletus sünohulgale „infotehnoloogia asjatundja, IT-spetsialist" oleks sel juhul „info automaatse töötlemise ja salvestamise asjatundja".

7% valimi sünohulkadele genereeritud seletused said hindeks 0 ehk ei sobinud (joonis 3). Peamiseks põhjuseks oli see, et seletus koosnes ainult sünonüümidest või seletus ei kirjeldanud sünohulka. Näiteks sünohulgale „depravatsioon, lagastamine, rikkumine" genereeriti seletus „halvenemine, pahenemine, raiskus, rikutus", mis koosneb ainult sõna „depravatsioon" sünonüümidest. Seletus, mis ei kirjeldanud sünohulka oli näiteks sünohulgale „meloodiaõpetus, meloodika" genereeritud seletus „klahvidega puhkpill". Selline seletus saadi, sest Ekilexi tehti päring sõnaga „meloodika" ning see on esimene seletus, mis sellele sõnale on antud. Lisaks anti hinne 0 ka seletustele, mis olid Ekilexis võõrkeelsed. Näiteks sünohulgale „medulloblastoom" genereeriti norrakeelne seletus „hjernesvulst i taket av fjerde ventrikkel og tilgrensende deler av lillehjernen". Lisaks norrakeelsetele seletustele esines ka venekeelne seletus.



Joonis 3. Unikaalse sünohulga liikme põhjal genereeritud seletuste hinnete jaotus

Kokkuvõttes saab sel viisil kõige paremini genereerida seletusi üheliikmelistele sünohulkadele. 78% hinde 4 saanud valimi seletustest genereeriti üheliikmelistele sünohulkadele ning 16% hinde 4 saanud seletustest genereeriti kaheliikmelistele sünohulkadele. Sellest võib järeldada, et mida rohkem liikmeid sünohulgas on, seda halvemini meetod töötab. Näiteks genereeriti üheliikmelisele sünohulgale „sigaretipaber" seletus „väga õhuke paber sigarettide kestade valmistamiseks". Lisaks ei saa unikaalse sünohulga liikme põhjal genereerida sõnaseletusi, milles esineb kirjavigu, sest kasutatakse Ekilexis esinevaid seletusi.

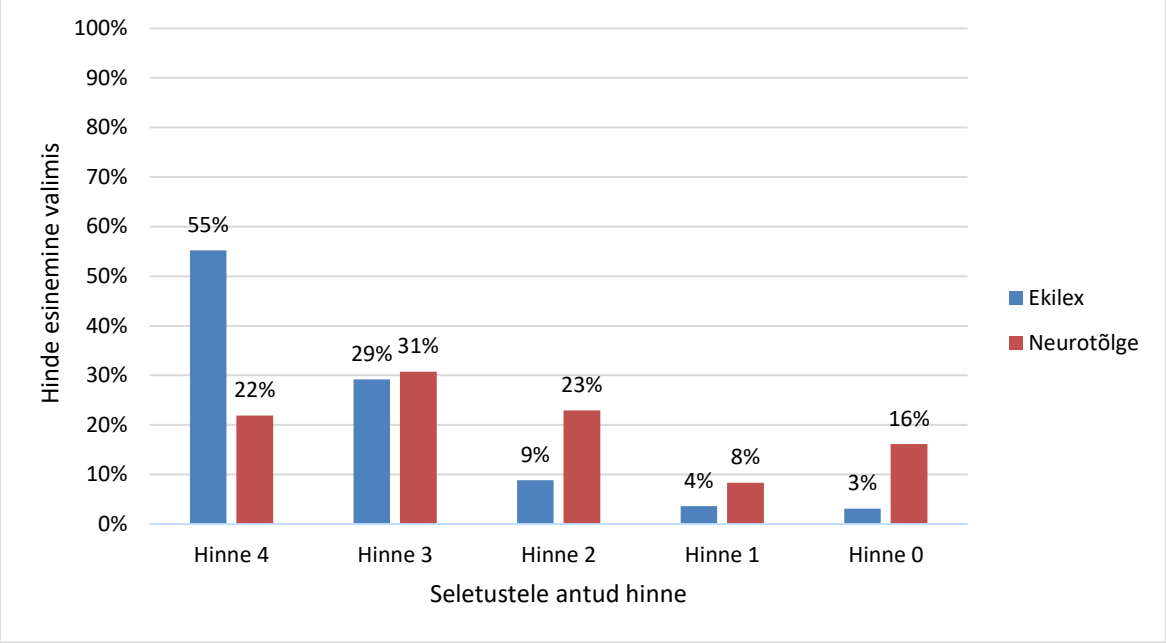
5.3 *Mine*-tegevusnime meetodiga genereeritud seletuste analüüs

Mine-tegevusnime meetodit kasutati 958 korral ning korrektsuse analüüsiks juhuslikult valitud valimi suurus oli 192 (joonis 1). Valimi keskmine seletuste hinne enne neurotõlke kasutamist oli 3,30 ning eesti keelest eesti keelde Tartu NLP Neurotõlkega tõlgitud seletuste keskmine hinne oli 2,34. Sobiv sõnaseletus genereeriti antud meetodiga valimis 55% juhtudest ning neurotõlke kasutamine seletust sageli paremaks ei muutnud (joonis 4). Näiteks sünohulgale „lätistamine" genereeriti algselt seletus „lätipäraseks muutmine" ning neurotõlge tõlkis seletust, kui „läti muutmine", mis on ebasobiv seletus. Neurotõlke kasutamine parandas mõnel korral kirjavigu, mis olid tekkinud peale *ma*-liite teisendamist *mine*-liiteks. Näiteks sünohulgale „altereerimine" genereeriti seletus „heli kõrguse muutmine", neurotõlge parandas seal kirjavea ning lõpptulemus oli „helikõrguse muutmine". Esines ka

olukordi, mil neurotõlge muutis sünohulga seletuse paremaks. Näiteks sünohulgale „eritumine" genereeriti algselt seletus „millestki välja eraldumine", kuid neurotõlge muutis selle „eraldumine millestki", mis on korrektsem, sest eraldumine ei pea olema alati millegi seest välja. Kokku teisendas neurotõlge valimis algselt saadud sõnaseletuse sobivaks seletuseks 20 korral.

Hinde 3 said sõnaseletused, mis kirjeldasid sünohulka hästi, aga esines üksikuid kirja- või lausestusviga. Näiteks sünohulgale „taageldamine, taglastamine" genereeriti seletus „purjelaevale paigaldamine trossidega süsteemi ehk taglase, mis kannab purjede, tulede, signaallippide vms", mis kirjeldab sünohulka hästi aga selles esineb kirjaviga ning oleks piisanud ka seletusest „purjelaevale trossidega süsteemi ehk taglase paigaldamine". Hinde 3 said kokku 29% Ekilexi info põhjal genereeritud seletustest ning 31% neurotõlke seletustest (joonis 4).

Hinde 0 said valimis ainult 3% Ekilexi info põhjal genereeritud seletustest ning neid sõnaseletusi ei muutnud neurotõlge paremaks (joonis 4). Näiteks sünohulgale „hammustamine, nähvamine, nipsamine, pistmine, salvamine, suskamine, torkamine" genereeriti algselt seletus „korraks kergelt löömine", sest Ekilexist tehti päring sõnaga „nipsama" ning teisendati selle seletus õigesse vormi. 16% neurotõlke seletustest said hinde 0 ning peamiseks veaks oli ühe ja sama sõna kordus (joonis 4). Näiteks sünohulgale „võnksumine" saadi algselt seletus „jõnksude tegemine, jõnksumine" ning neurotõlge muutis selle „tõukamine, tõukamine". Kuid leidus ka seletusi, kus neurotõlge teisendas lause täiesti teisele kujule. Näiteks sünohulgale „kühveldamine" saadi Ekilexist seletus „kühvliga, ka labidaga või muu sobiva riistaga millegi teise kohta tõstmine või pildumine". Neurotõlge tõlkis vastavat seletust, kui „tõste- või teisaldus mehhanismiga varustatud iseliikuvad veokid", mis on selgelt vale.



Joonis 4. *Mine*-tegevusnime meetodiga genereeritud sõnaseletuste hinnete jaotus

Analüüsi tulemustest võib järeldada, et eesti keelest eesti keelde tõlkimine enamasti sõnaseletust paremaks ei tee. Neurotõlke kasutamine valimi seletustel parandas hinnet ainult 28 korral. Esines ka juhte, kus neurotõlge ei muutnud sisestatud seletust. Peamised meetodi kasutamisel esinevad vead olid kirja- ning lausestusvead, mille harva parandas neurotõlke kasutamine. Kokkuvõttes on võimalik genereerida antud meetodiga sobivaid seletusi, kuid enne seletuse lisamist EstWNi tuleb see kindlasti üle kontrollida, sest peaaegu pooltel kordadel on seletus ebasobiv või esineb seletuses viga.

5.4 *Ja*-tegijanime meetodiga genereeritud seletuste analüüs

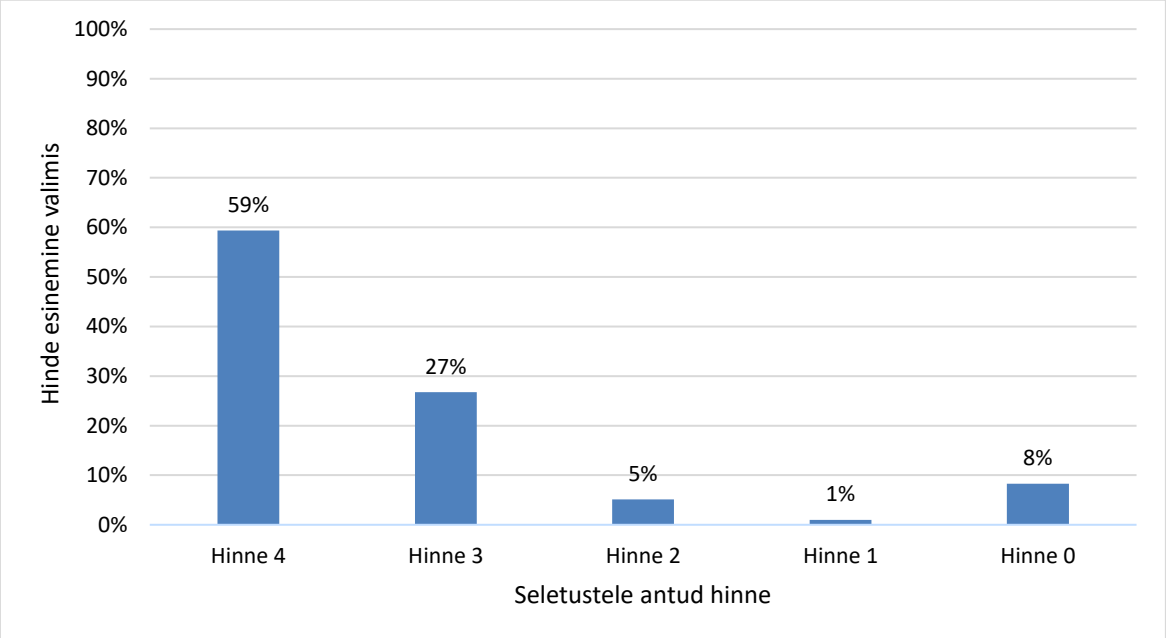
Ja-tegijanime meetodit kasutati sõnaseletuse genereerimiseks 2053 korral ning analüüsitava valimi suurus oli 411 sõnaseletust (joonis 1). Valimi seletuste keskmiseks hindeks oli 3,28. Hinde 4 saanud seletusi oli valimis kokku 59% (joonis 5). Näiteks hinde 4 sai sünohulgale „avalikustaja" genereeritud seletus „keegi, kes avalikuks teeb". Genereeritud seletus on piisava pikkusega, ilma kirja-või lausestusvigadeta ning kirjeldab sünohulka piisavalt hästi.

Hinde 3 said enamasti seletused, mis sisaldasid mõnda kirja- või lausestusviga Põhiliseks veaks oli see, et kõik seletused genereeriti tegijatele, kes on elusolendid, kuigi tegijad võivad olla ka mitte elusolendid. Näiteks sünohulgale „puhverdaja" genereeriti seletus „keegi, kes infot ajutiselt salvestab, seda vastavasse mäluossa, puhvrisse, paigutab", mis viitab puhverdajale kui elusolendile. Saadud sõnaseletus sobiks juhul, kui asendada „keegi, kes" väljendiga „miski, mis". Esines ka juhte, kus tegija võis olla nii elusolend kui ka mitte elus-

olend. Sellisel juhul tuleks asendada „keegi, kes" väljendiga „keegi või miski, mis". Näiteks sünohulgale „ergastaja" genereeriti seletus „keegi, kes organismi (vaimult, kehalt) virgemaks, erksamaks teeb", kuid „ergastaja" võib olla ka antud juhul „miski, mis", mitte ainult „keegi, kes".

Seletused, kus esines nii kirja- või lausestusvigu ning ei kirjeldanud kõige paremini sünohulka said hindeks 2. Hinde 2 saanud sünohulki oli valimist 5% (joonis 5). Näiteks sünohulgale „lõgiseja, plõgiseja" genereeriti seletus „keegi, kes üksteisele tihedalt järgnevaid metalseid helisid kuuldavale laseb", mis on halvasti sõnastatud ning ei kirjelda vastavat sünohulka hästi, sest lõgina või plõgina heli ei pruugi olla metalne. Parem seletus oleks „miski, mis üksteisele tihedalt järgnevaid kergelt kõlavaid helisid kuuldavale laseb".

Ainult 8% valimi genereeritud sõnaseletustest ei sobinud (joonis 5). Peamiseks põhjuseks oli, et Ekilexist saadud seletus oli liiga spetsiifiline sellele sõnale, millega päring tehti. Seetõttu ei kirjeldanud saadud seletus kogu sünohulka piisavalt hästi. Näiteks sünohulgale „isemeelitseja, kapriisitseja, pahurdaja, tujutseja, tuuritaja" genereeriti seletus „keegi, kes oma tahtmist mööda, isemeelselt käitub", mis aga kirjeldab ainult sõna „isemeelitseja". Saadud seletus saadi seetõttu, et Ekilexi tehti päring sõnaga „isemeelitsema". Siinkohal oleks seletust parandanud sünohulga liikmete muutmine, teha näiteks mitu sünohulka.



Joonis 5. *Ja*-tegijanime meetodiga genereeritud sõnaseletuste hinnete jaotus

Kokkuvõttes saab sel viisil *ja*-liitega sünohulkadele genereerida sobivaid seletusi, kuid sõnaseletused tuleks enne EstWNi lisamist üle kontrollida, sest 41% valimi seletustest

sisaldasid endas kirja- ja/või lausestusvigu või ei sobinud sünohulga seletuseks. Peamiseks veaks osutus „keegi, kes" seletuse algus, sest tegijad võivad olla ka mitte elusolendid („miski, mis") või siis mõlemat („keegi või miski, mis"). Siiski võib valimi seletustest tuletada, et enamasti on tegijad elusolendid, sest üle poolte genereeritud seletustest sobis.

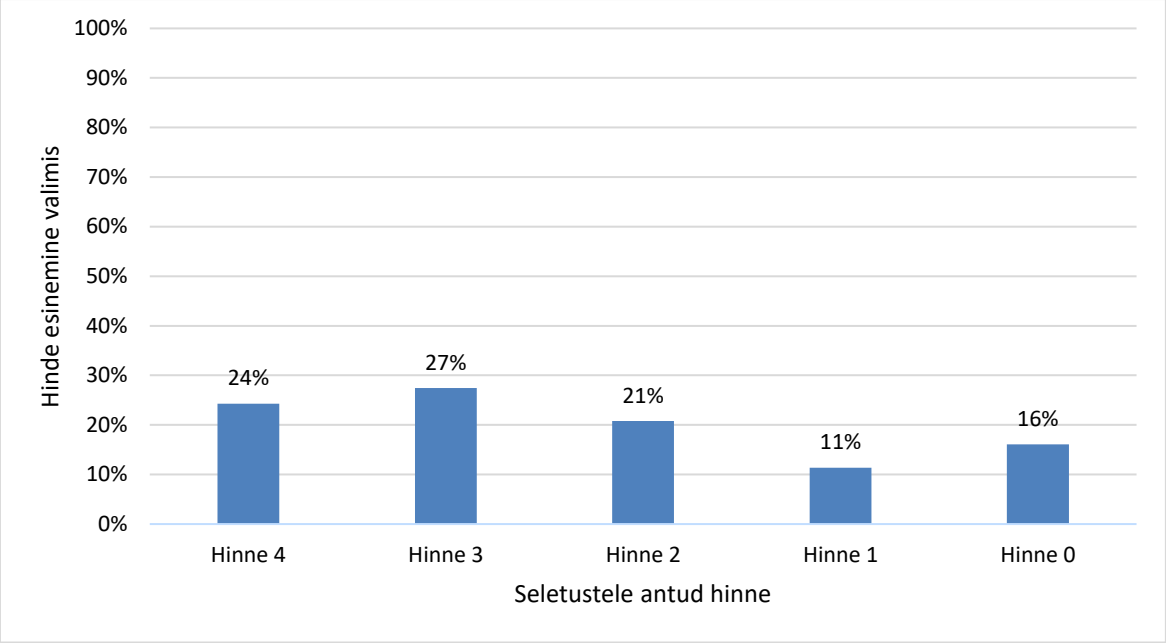
5.5 Tõlgitud seletuste analüüs

Ainult Princetoni Wordnetist saadud sõnaseletusi oli kokku 1277 ning analüüsitava valimi suuruseks oli 255 (joonis 1). Tõlgitud sõnaseletuste keskmiseks hindeks valimis oli 2,33. Sel viisil saadi valimis kõige vähem sobivaid sõnaseletusi ning ainuke meetod, mille korral hinne 4 ei olnud kõige populaarsem (joonis 6). Hinde 4 sai näiteks sünohulgale „nina-kitsenemus, rinostenoos" genereeritud seletus „ninaõõne läbipääsude ahenemine", sest saadud seletuses ei esine kirja- ega lausestusvigu ning kirjeldab sünohulka piisavalt täpselt ja arusaadavalt lahti.

Hinde 3 said 27% valimi tõlgitud seletustest (joonis 6). Peamiseks põhjuseks oli, et üks sõna tõlgiti valesti, aga seletus andis sünohulga tähenduse piisavalt hästi edasi. Näiteks sünohulgale „fundamentaalsus, põhjalikkus, põhjapanevus" saadi seletus „mis tahes tegur, mida võib pidada oluliseks konkreetse ettevõtte mõistmisel". Saadud seletus sobib juhul kui asendada sõna „ettevõtte" sõnaga „valdkonna".

Hinde 2 või 1 said seletused, mis ei kirjelda sünohulka piisavalt hästi või/ja kus esines kirja- või lausestusvigu. Näiteks sünohulgale „paksendamine, paksendus" saadi seletus „paksenemise tegu", mis aga on liiga üldine ning liiga lühikene seletus. Kokku sai valimist 21% tõlgitud seletustest hinde 2 ja 11% tõlgitud seletustest hinde 1 (joonis 6).

Võrreldes teiste meetoditega oli tõlgitud seletuste seas kõige rohkem seletusi, mis ei sobinud ehk said hindeks 0. Hinde 0 said kokku 16% valimi seletustest (joonis 6). Näiteks sünohulgale „tabamatus" genereeriti seletus „raskesti hoomatava või allasurutava kvaliteedi", mis on halvasti sõnastatud ning ei anna kuidagi edasi sünohulga tähendust.

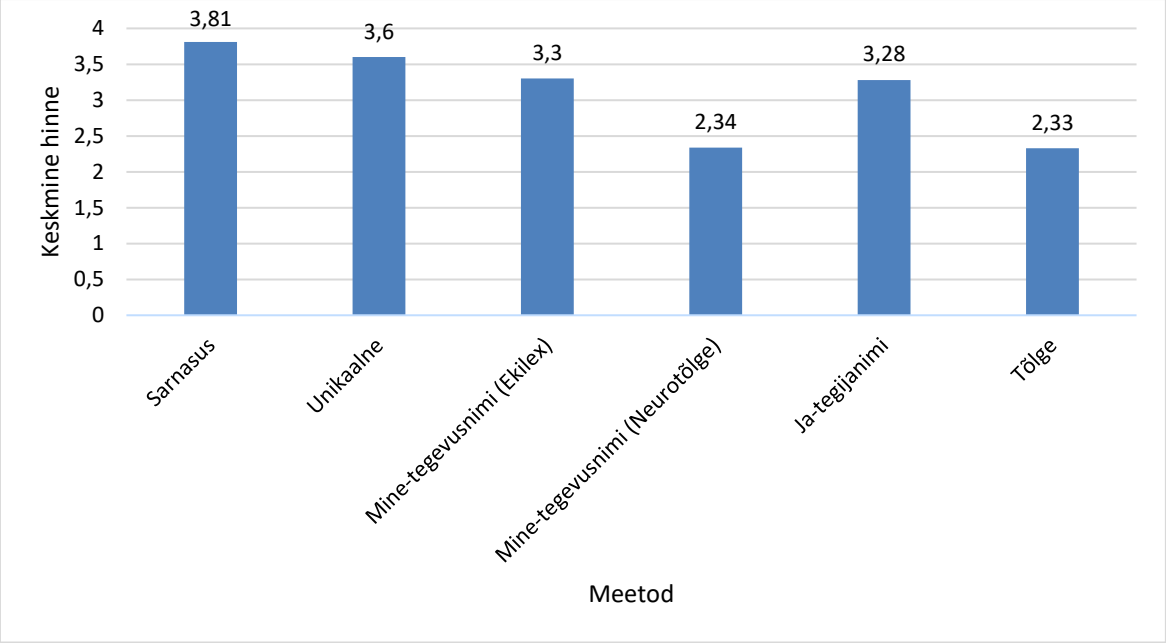


Joonis 6. Tõlgitud sõnaseletuste jaotus

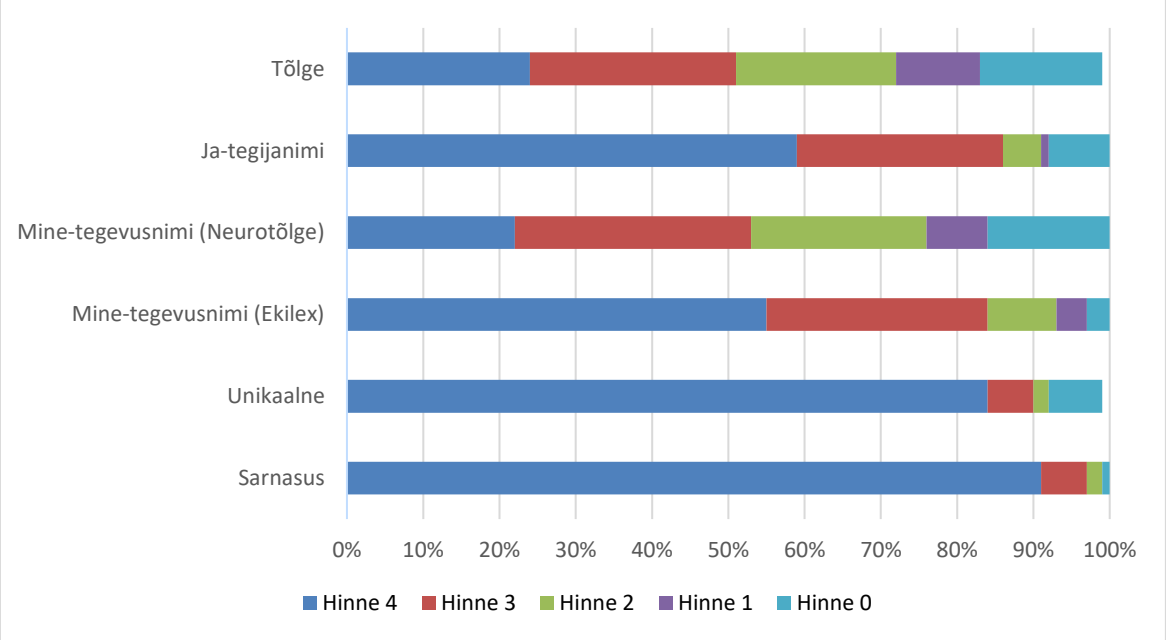
Kokkuvõttes sobisid inglise keelest eesti keelde tõlgitud seletused vaid veerand juhtudel. Palju suurema tõenäosusega on TartuNLP Neurotõlkega saadud tulemus vigane või mittesobiv. Vigastes seletustes oli tihtipeale mõni sõna üleliigne, esines sõna kordust või kirja- ja lausestusvigu. Seetõttu tuleks sel viisil saadud seletused alati enne EstWNi lisamist üle kontrollida.

5.6 Analüüsi kokkuvõte

Kõige kindlam viis sobiva sõnaseletuse genereerimiseks, nii keskmise hinde (joonis 7) kui ka hinnete protsentuaalse jaotuse (joonis 8) alusel, on kasutada sarnasuse ja unikaalse sünohulga liikme meetodit. Üle poolte *mine*-tegevusnime ja *ja*-tegijanime meetodiga genereeritud valimi sõnaseletustest olid küll sobivad, aga üle 40% valimi seletustest sisaldasid vigu või olid mõnel muul põhjusel ebasobivad (joonis 8). Princetoni Wordnetist tõlgitud seletuste tulemus oli teiste meetoditega võrreldes kõige halvem, sest tõlgitud seletus ei kirjeldanud vastavat sünohulka. Neurotõlkega (est→est, eng→est) saadud seletused pigem ei sobinud, sest need ei kirjeldanud vastavat sünohulka.



Joonis 7. Meetodite keskmised hinded



Joonis 8. Hinnete protsentuaalne jaotus erinevate meetodite puhul

Mine-tegevusnime ning *ja*-tegijanime meetoditega genereeritud seletuste peamisteks vigadeks osutusid kirja- ja lausestusvead. *Mine*-tegevusnimede puhul aitas neurotõlke kasutamine neid mõnel juhul parandada, kuid enamasti neurotõlke kasutamine seletust paremaks ei muutnud. Sarnasuse põhjal ja unikaalse sünohulga liikmega saadud seletuste kõige sagedasemateks vigadeks osutus nende hägusus ehk ei seletanud sünohulka piisavalt hästi või siis ei seletanud kogu sünohulka. Esines ka seletusi, mis olid liiga spetsiifilised ja pikad.

Saadud valimi seletuste seas esines ka seaduspära, et mida suurem sünohulk on, seda väiksem tõenäosus on saada nende meetoditega vastavale sünohulgale sobiv seletus. Kokkuvõttes, kuna ükski meetod valimi korral 100% täpsusega ei töötanud, tuleks iga seletus enne Eesti Wordneti lisamist üle kontrollida. Lisaks genereeriti nende meetoditega sobivaid sõnaseletusi omadussõnadest koosnevatele sünohulkadele väga vähe, sest omadus-sõnad on Ekilexis peamiselt defineeritud sünonüümide kaudu, mis EstWNis moodustavad sünohulga.

6. Kokkuvõte

Selle bakalaureusetöö eesmärk oli automaatselt genereerida ja kontrollida seletusi Eesti Wordneti sünohulkadele, millel seletus puudus. Töö teoreetiline osa andis ülevaate arvuti leksikonide ajaloost, *wordnet*-tüüpi leksikonide põhimõtetest ning nende erinevatest loomisviisidest. Peamiselt keskendus teoreetiline osa Eesti Wordneti ning selle probleemide tutvustamisele. Lisaks seletati, kuidas koostada ja formaliseerida sõnaseletust.

Töö raames koostatud rakendus genereeris EstWNi sünohulkadele seletusi neljal erineval viisil: sünonüümide sarnasuse, unikaalse sünohulga liikme, *mine*-tegevusnime algvormi ning *ja*-tegijanime algvormi põhjal. Lisaks kontrolliti iga sünohulga puhul, kas Princetoni Wordnetis leidub võrdne sünohulk. Juhul kui leidus, kasutati Princetoni Wordnetist saadud sünohulga seletuse tõlkimiseks TartuNLP Neurotõlget ning lisati lõpp-tulemusele juurde. *Mine*-tegevusnime meetodi poolt genereeritud seletuste puhul kasutati lisaks TartuNLP Neurotõlke eesti keelest eesti keelde tõlkimist, et kontrollida, kas sel viisil saab seletuse paremaks muuta.

Kokku genereeris rakendus 11 075 sõnaseletust 18 731 puuduvast Wordneti seletusest. 7656 sünohulgale nende meetoditega seletust ei saadud. Kõige tulemuslikum meetod oli sünohulga seletuse leidmine unikaalse sünohulga liikme põhjal, millega genereeriti 5469 sõnaseletust. Kõige vähem sõnaseletusi genereeriti *mine*-tegevusnime meetodiga – 958 sõnaseletust. Iga meetodi analüüsiks valiti 20% selle meetodi poolt genereeritud seletustest ning valimi seletusi hinnati skaalal nullist neljani, kus neli on sünohulgale sobiv seletus ja null on mittesobiv seletus. Analüüsist võib järeldada, et kõige kindlam viis sõnaseletuste genereerimiseks on kasutada sarnasuse meetodit. Sarnasuse meetodi põhjal genereeriti sobiv seletus 91% valimi juhtudest. Kõige halvemini töötas Princetoni Wordneti seletuse tõlkimine eesti keelde. Lisaks parandas TartuNLP Neurotõlke eesti keelest eesti keelde tõlkimine *mine*-tegevusnime meetodiga saadud seletused väga harva paremaks. Analüüsitud seletustes väljendus seaduspära – mida suurem sünohulk, seda väiksem on tõenäosus nende meetoditega vastavale sünohulgale sobiv seletus genereerida.

Viidatud kirjandus

Atkins, B. T. ja Rundell, M. (2008). Building the monolingual entry. *The Oxford Guide to Practical Lexicography* (pp. 385–465). Oxford: Oxford University Press. Vaadatud 03.04.2022 https://nubip.edu.ua/sites/default/files/u5/atkins_b_t_s_rundell_m_the_oxford_guide_to_practical_lexicog.pdf

Eesti Wordneti koduleheküIg. (i.a). Vaadatud 06.12.2021 https://www.cl.ut.ee/ressursid/teksaurus/?lang=et

EKI sõnastiku- ja terminibaasisüsteem (Ekilex). Vaadatud 10.12.2021 https://ekilex.ee/

EstNLTK dokumentatsioon. (i.a). Vaadatud 12.02.2022 https://estnltk.github.io/

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Global WordNet Association koduleheküIg. (i.a). Vaadatud 31.03.2022 http://globalwordnet.org/resources/wordnets-in-the-world/

EKI ühendsõnastik 2022. Eesti Keele Instituutt, Sõnaveeb 2022. Vaadatud 10.04.2022 https://sonaveeb.ee/search/unif/dlall/dsall/ilus/1

Kahusk, N. ja Vider, K. (2017). The Revision History of Estonian Wordnet. *Proceedings of the LDK 2017 Workshops: 1st Workshop on the OntoLex Model (OntoLex-2017), Shared Task on Translation Inference Across Dictionaries & Challenges for Wordnets co-located with 1st Conference on Language, Data and Knowledge (LDK 2017)*. Galway, June 18 (pp. 164–173). CEUR-WS.org: Creative Commons. Vaadatud 15.11.2021 http://ceur-ws.org/Vol-1899/CfWNs_2017_proc5-paper_6.pdf

Kilgarriff, A. (2000). Review: WordNet: An Electronic Lexical Database. *Language, 76 (3),* 706–708. Vaadatud 07.12.2021 https://www.jstor.org/stable/417141?origin=crossref

Meos, I. (2003). Mõiste. *Loogika. Argumentatsioon. Mõtlemiskultuur* (lk 13–26). Tallinn: Koolibri. Vaadatud 03.04.2022 https://core.ac.uk/download/pdf/14484624.pdf

Muischnek, K., Orav, H., Kaalep, H.-J. ja Õim, H. (2003). Ülevaade arvutileksikonide arengust. U. Talvik (toim), *Eesti keele tehnoloogilised ressursid ja vahendid : arvutikorpused, arvutisõnastikud, keeletehnoloogiline tarkvara* (lk 27–29). Tallinn: Eesti Keele Sihtasutus.

Nesi, H. (2009). Dictionaries in electronic form. A.P. Cowie (ed.), *The Oxford History of English Lexicography* (pp. 458–478). Oxford: Oxford University Press. Vaadatud 09.12.2021 https://www.researchgate.net/publication/267777639_Dictionaries_in_electronic_form

Orav, H., Vare, K. ja Zupping, S. (2015). Leksikosemantiliste suhete hägusus Eesti Wordnetis. M. Erelt (toim) ja M.-M. Sepper(toim), *Emakeele seltsi aastaraamat 60* (lk 171–194). Tallinn: Emakeele Selts. Vaadatud 14.01.2022 https://www.kirj.ee/public/ESA/2014/esa_60_2014_171-194.pdf

Orav, H., Kerner, K. ja Parm, S. (2011). Eesti Wordnet'i hetkeseisust. *Keel ja Kirjandus, 2*, 96–106. Vaadatud 03.03.2022 https://keeljakirjandus.eki.ee/96-106.pdf

Pedersen, B. S., Borin, L., Forsberg, M., Kahusk, N., Lindén, K., Niemi, J., Nisbeth, N., Nygaard, L., Orav, H., Rögnvaldsson, E., Seaton, M.,Vider, K. ja Voionmaa, K. (2013). Nordic and Baltic wordnets aligned and compared through "WordTies". *Proceedings of the 19th Nordic Conference of Computational Linguistics (NODALIDA 2013)*. Oslo, May 22–24 (pp. 147–162). Linköping University Electronic Press: Linköping University. Vaadatud 01.04.2022 https://ep.liu.se/ecp/085/016/ecp1385016.pdf

Princetoni Wordneti koduleheküIg. (i.a). Vaadatud 31.03.2022 https://wordnet.princeton.edu/

TartuNLP Neurotõlge. (i.a). Vaadatud 14.02.2022 https://translate.ut.ee/

Teder, T. ja Viikberg, J. (2017). Eesti keele instituut. Algusaastad. *Oma keel, 34(1)*, 73–79. Vaadatud 06.01.2022 https://www.emakeeleselts.ee/omakeel/2017_1/OK-1-2017_12.pdf

The history of COBUILD. Collins Dictionary Blog. Vaadatud 09.12.2021 https://blog.collinsdictionary.com/the-history-of-cobuild/0/

Uiboaed, K. (2005). *Sõnaseletuse genereerimine tesauruse info põhjal*. Bakalaureusetöö. Tartu Ülikool, arvutilingvistika õppekava.

Villem, O.-A. (2009). *ILI-kirjete lisamine eesti wordnetti ja selle käigus ilmnenud automaatselt genereeritud sünohulkade probleemkohad*. Bakalaureusetöö. Tartu Ülikool, eesti ja üldkeeleteaduse instituut.

Lisad

I. Litsents

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Kristo Markov

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Sõnastike automaatne genereerimine ja kontrollimine Eesti Wordneti näitel“**,** mille juhendajad on Heili Orav ja Indrek Jentson 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.

Kristo Markov

10.05.2022