Tartu Ülikool
eesti ja üldkeeleteaduse instituut
eesti ja soome-ugri keeleteadus
arvutilingvistika eriala

Eesti Wordnet ja meelestatuse analüüs

Gerth Jaanimäe

Magistritöö

Juhendaja: Heili Orav, Ph.D.

Tartu
2018

Sisukord

Sissejuhatus

Meelestatuse analüüs on tänapäeva maailmas üsna oluline uurimisvaldkond. Inimesed avaldavad oma arvamusi sotsiaalmeedias, blogides, foorumites, e-poodide kasutajaülevaadetes, internetinetikommentaarides ja mujal. Kuna ettevõtete jaoks on oluline teada, mida kasutajad nende toodetest arvavad, oleks väga oluline, kui saaks automaatselt kindlaks teha, mis inimestele meeldib ja mis mitte. Meelestatuse analüüsil on ka teisi kasutusvaldkondi, nagu näiteks terrorismi vastu võitlemine.

Käesoleva magistritöö eesmärk on luua uus keeleressurss, kus nii Eesti Wordnet kui meelestatuse skoorid on omavahel automaatselt ühendatud. Eesti Wordnetist on tänu oma suurele leksikonile praeguseks kujunenud arvestatav keeleressurss ning näiteks inglise ja poola keele *wordnet*'iga on katsetatud mitmeid erinevaid lähenemisi meelestatuse märgendite lisamiseks.
Teiseks eesmärgiks on uurida, millise kvaliteediga tulemus saadakse, kui võetakse automaatselt üle teise keele andmestik, st kas ja kui hästi see meelestatuse analüüsiks eesti keele puhul toimib.
Kolmas eesmärk on hinnata, kas kõik sünonüümid kannavad enda tähenduses sama meelestatust, sest *wordnet*-tüüpi sõnastiku oluline osa on just sünonüümsete sõnade hulgad.

Töö koosneb järgmistest osadest.
Teoreetilises osas tutvustatakse kõigepealt meelestatuse analüüsi meetodeid üldiselt ning kirjeldatakse lühidalt nende eeliseid ja puudusi. Järgnevalt tutvustatakse Eesti Wordneti aluspõhimõtteid. Seejärel kirjeldatakse erinevaid *wordneti*-põhiseid lähenemisi meelestatuse analüüsiks. Käesolevas magistritöös keskendutakse SentiWordNetile ja SenticNetile.
Praktilises osas kirjeldatakse alustuseks töö käiku, loodud skriptide tööpõhimõtteid ning töö tulemusi. Seejärel tutvustatakse magistritöö käigus läbiviidud pilootuuringut ning sellest tehtud järeldusi.
Töö lõpuosas tutvustatakse potentsiaalseid edasiarendusvõimalusi ning arutletakse *wordneti* ja meelestatuse analüüsi teemal üldisemalt. Loodud skriptid ja valminud andmebaas on ära toodud lisas.

1. Taustast

Inimeste otsustusi mõjutavad alati teatud määral teiste mõtlemine, ideed ja arvamused. Suurenev sotsiaalmeedia osatähtsuse kasv suurendab selle kasutajate avaldatud kommentaaride, ülevaadete ja arvamuste hulka mingi toote, teenuse või ürituse kohta. Sellised tekstid on kasulikud nii tarbijatele kui ka tootjatele. Esimesed soovivad enne ostu sooritamist teada, mida teised sellest arvavad. Tootjad saavad jällegi aimu nende toodete tugevatest ja nõrkadest külgedest. Kuna aga selliseid andmeid on väga suur hulk, muutub see aga nende lugejaile raskesti hallatavaks. Seega küsimus, kuidas sellist suurt tekstiandmete hulka analüüsida ja kokku võtta, on uurijate jaoks väga huvitav valdkond.

Meelestatuse analüüs (ingl *sentiment analysis* või *opinion mining*) on automaatne suhtumiste, arvamuste ja emotsioonide tuvastamine tekstidest, kõnest või muud tüüpi andmetest loomuliku keele töötluse kaudu. Meelestatuse analüüsi protsessi käigus klassifitseeritakse arvamused tavaliselt kas positiivseks, negatiivseks või neutraalseks ja seda saab teha mitmel erineval viisil. (Vohra, Teraiya 2013)

1.1 Meelestatuse analüüsi erinevad lähenemised

Meelestatuse analüüsimiseks on siiani kasutusel olnud peamiselt kaks erinevat lähenemist: esiteks masinõppel ja teiseks leksikonil põhinev lähenemine. Järgnevas antakse neist ülevaade, tuginedes S. M. Vohra ja J. B. Teraiya artiklile "A Comparative Study of Sentiment Analysis Techniques" (2013).

Esimene lähenemine kasutab põhiliselt juhendatud klassifitseerimisalgoritme. Selleks jaotatakse suur hulk tekste kaheks: treening- ja testandmeteks. Treeningandmete pealt õpib algoritm tekste üksteisest erinevate tunnuste põhjal eristama. Testandmetega kontrollitakse, kui hästi treenitud algoritm toimib. Tunnuseid, mille alusel klassifitseerimist õpetatakse, võivad olla n-grammide (sõnade, sõnapaaride või -kolmikute) sagedused, sõnaliigi kohta käiv info jms.

Masinõppepõhise lähenemise peamiseks eeliseks on suurem täpsus mingi kindla valdkonna tekstide peal, miinuseks aga see, et efektiivseks kasutamiseks läheb vaja suurel hulgal treeningandmeid.

Tuntumad masinõppealgoritmid, mida meelestatuse määramiseks kasutatakse on tugivektormasin (ingl *support vector machine*), maksimum entroopia (ingl *maximum entropy*)

ja Naive Bayes. Esimene neist töötab hästi (st määratleb meelestatuse täpsemini) juhul, kui on palju andmeid, väikese andmehulga puhul toimib aga hästi viimane nimetatuist.

Leksikonil põhineva lähenemise puhul võrreldakse tekstis olevaid sõnu juba eelnevalt koostatud leksikonis olevatega. Sõnastik sisaldab sõnu ja väljendeid, millele on lisatud meelestatuse märgendid või skoorid.
Teksti analüüsimisel vaadatakse, kas see sisaldab rohkem positiivse või negatiivse meelestatusega sõnu. Esimesel juhul loetakse kogu tekst positiivseks, viimasel juhul aga negatiivseks. Tekstide klassifitseerimiseks pole vaja treeningandmeid, sellepärast nimetatakse sellist lähenemist juhendamata meetodiks.

Analüüsimiseks kasutatakse enamasti järgmist protsessi:

1. Teksti eeltöötlus – eemaldatakse HTML-märgendid ning muud erisümbolid.
2. Algväärtustatakse teksti meelestatuse skoor nulliga.
3. Teksti sõnestamine. Iga sõne puhul kontrollitakse, kas see meelestatuse leksikonis eksisteerib. Kui jah ning sõne on positiivse märgendiga, liidetakse vastav skoor kogu teksti omale. Kui negatiivne, siis see lahutatakse.
4. Vaadatakse teksti kogu skoori. Kui see on algväärtusest suurem, klassifitseeritakse tekst positiivseks, kui väiksem, siis negatiivseks.

Meelestatuse leksikoni koostamiseks on kolm erinevat võimalust: käsitsi, korpusepõhine ja sõnastikupõhine.
Käsitsi leksikoni koostamine on teadaolevalt pikk ja vaevarikas protsess, kus kulub palju (inim)ressurssi. Sõnastiku põhise leksikoni koostamisel võetakse väike hulk sõnu, milles meelestatus on juba ära määratletud ning suurendatakse seda hulka näiteks *wordnet*'ist sünonüüme ja antonüüme võttes. Selle meetodi puudus seisneb peamiselt selles, et see ei tööta väga hästi spetsiifiliste valdkondade tekstide peal.

Korpustel põhinev lähenemine toetub suurtes tekstikorpustes olevatele süntaktilistele mustritele. Selle eeliseks on üsna hea täpsus, ka spetsiifiliste valdkondade tekstide puhul. Puudus on aga sama, mis masinõppepõhiste lähenemiste puhul, ehk siis vaja on suurt andmehulka.

Katsetatud on ka hübriidlähenemist, kombineerides omavahel masinõppel ja leksikonil põhinevaid meetodeid. Peamiseks eeliseks sel puhul on mõlemate meetodite positiivsete külgede ühendamine. Leksikaalse lähenemise stabiilsus ja hästi disainitud leksikoni kerge loetavus ning masinõppepõhise lähenemise täpsus. (Vohra, Teraiya 2013)

Käesoleva magistritöö jaoks valiti leksikonipõhine lähenemine, sest Eesti Wordnet on suhteliselt mahukas ja sarnast lähenemist on edukalt katsetatud ka inglise keele puhul. Magistritöö on esimene katse koostada automaatselt uus meelestatuse infot sisaldav ressurss eesti keele jaoks.

1.2. Eesti Wordnetist

Eelmise sajandi 80-ndatel aastatel loodi Princeton WordNet[1], peale seda sai alguse ka teiste keelte leksikosemantiliste andmebaaside loomine, mida kutsutakse üldnimega *wordnet*. Järgnevalt antakse ülevaade *wordnet*-tüüpi sõnastiku ülesehitusest ning konkreetsemalt Eesti Wordnetist tuginedes peamiselt artiklile "Leksikosemantiliste suhete hägusus Eesti Wordnetis" (Orav jt 2014).

Selle ideeline põhi on leksikaalsete üksuste võrgustik, kus seotus tuleneb teatud fikseeritud suhete valikust. Algidee oli luua sõnavõrgustiku tüüpi mentaalne leksikon ehk mudel selle kohta, kuidas sõnad meie peas asetuvad ja kuidas need omavahel seotud on. Kirjeldatud leksikon oli mõeldud esmalt psühholoogide ja keeleteadlaste jaoks, nende uurimistulemuste peegeldamiseks, kuid praeguseks on *wordnet* pigem keeletehnoloogide tähelepanu all olev ressurss. *Wordneti* kui väärtusliku keeleressursi tõusmine infotehnoloogia valdkonda on põhjustatud eelkõige vajadusest selgitada arvutisüsteemidele loomuliku keele mõisteseoseid, st arvuti peaks keeleandmete põhjal oskama ka teatud semantilisi järeldusi teha (nt mets koosneb puudest, pahtel on teatud ehitusmaterjal). *Wordnet*'ide eeliseks paljude teiste sõnastike ees on mitmekeelsus – eri keelte *wordnet*'id on omavahel ühendatud keeltevahelise indeksiga, mis võimaldab mõistepõhiselt tõlkevasteid leida. Kõige esimesele ehk ingliskeelsele *wordnet*'ile on viimaste aastate jooksul tulnud lisa üle seitsmekümne keele kohta, sh näiteks ka surnud ladina keele wordnet.

Eesti Wordnet (EstWN) on tänapäevane leksikaalsemantiline andmebaas, mida kasutatakse mitmetes keeletehnoloogilistes rakendustes ja selle maht 2018. aasta alguses on üle 86 200 sünohulga (erinevaid sõnu üle 139 000)[2]. Sõnaliikidelt koosneb Eesti Wordnet adjektiividest, substantiividest, verbidest ja adverbidest, mis iga sõnaliigi sees on koondatud paljudesse tähenduslikesse üksustesse ehk sünohulkadesse (ingl *synset*) vaikimisi täis- ja lähisünonüümia suhte alusel moodustatud. Võrdluseks võib tuua maailma suurimad *wordnet*'id: Princetoni

[1] Vt https://wordnet.princeton.edu/ (18.05.2018)
[2] Andmed pärinevad https://www.keeletehnoloogia.ee/et/ekt-projektid/eesti-wordneti-taiendamine-2/eesti-wordneti-taiendamine-2 (18.05.2018)

WordNetis 3.1 on 175 979 sünohulka[3] ja poola keele wordnetis on sünohulkadesse jagatud 191 000 sõna[4].

Võrreldes ingliskeelse Princetoni WordNetiga on Eesti Wordnetis märksa rohkem eri tüüpi semantilisi suhteid, et veelgi täpsemalt anda edasi tähendusnüansse. Kuna EstWN-is on olemas võimalus siduda ka eri sõnaliikidest lähtuvaid mõisteid, siis moodustub neist mõistetest mingi konkreetne semantiline väli, valdkond – semantiliselt seotud sõnade hulk, mis moodustab teatud mõistelise terviku. Tänapäeva keeletehnoloogilised rakendused töötavad paremini paljuski just valdkondliku lähenemisega: näiteks sõnatähenduste ühestamine, masintõlge või infootsing saavad tänu tihedale semantilisele võrgustikule palju rohkem materjali, millega töötada. Samuti on *wordnet* oluline keeleteaduses näiteks keele leksikaalse struktuuri uurimisel, keeletehnoloogias tekstide automaatse kokkuvõtte tegemisel, sõnavaliku vigade automaatsel parandamisel tekstis jm. (Orav, Zupping, Vare 2014) EstWN-i suurendamise ja kvaliteedi parandamisega tegeletakse jätkuvalt, seega pole tegemist lõpetatud ja valmis ressursiga.

Edasi käsitletakse erinevaid ressursse, kus on meelestatuse info leksikonis olemas. Ehkki magistritöösse kasutati *wordneti*-põhist leksikoni, oli eelnevalt vaja tutvuda ja hinnata teisi ressursse ning lähenemisi meelestatuse analüüsiks. Järgnevalt neid tutvustataksegi.

1.3. SentiWordNet

SentiWordNet[5] on Princetoni Wordnetis olevate sünohulkade automaatse märgenduse tulemus, kus igale sünohulgale vastab kolm skoori: positiivsus, negatiivsus ja neutraalsus. Näiteks ingliskeelsel sünohulgal "estimable" on positiivsus 0,75, neutraalsus 0,25 ja negatiivsus 0,0. Iga skoor on vahemikus 0,0-1,0 ja nende summa võrdub alati ühega. Skoorid võivad olla ka kõik nullist erinevad, mis tähendab, et mõisted võivad vastata teatud määral mitmele meelsuse mõõdikule.

SentiWordNet 3.0 on loodud kahes etapis: pooljuhendatud masinõpe ja juhuslik liikumine (*random walk process*). Järgnev toetub Stefano Baccianella, Andrea Esuli, ja Fabrizio Sebastiani artiklile “SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining” (2010).

[3] Andmed pärinevad https://en.wikipedia.org/wiki/WordNet (18.05.2018)
[4] Andmed pärinevad http://plwordnet.pwr.wroc.pl/wordnet/ (18.05.2018)
[5] Vt http://SentiWordNet.isti.cnr.it/ (18.05.2018)

1.3.1. Pooljuhendatud masinõpe

Pooljuhendatud masinõppe etapp koosneb omakorda neljast sammust:
1) Andmestiku laiendamine
2) Klassifitseerija treenimine
3) Sünohulkade klassifitseerimine
4) Klassifitseerijate kombineerimine.

Esimeses sammus valitakse väikseks andmestikuks seitse paradigmaatiliselt positiivset ja seitse negatiivset sünohulka, mille abil hakatakse automaatselt andmestikku laiendama. Seda saab teha liikudes erinevate semantiliste suhete kaudu mööda *wordneti* sünohulki. Laiendust saab teostada teatud kindlas raadiuses, ehk siis arvestades sünohulkade omavahelisi kaugusi algsetes andmehulkades.
Teises sammus - klassifitseerija treenimises - kasutatakse klassifitseerija loomisel lisaks eelmises loodud andmestikule ka hulka, milles eeldatakse olevat neutraalsete mõistetega sünohulgad. Lisaks kasutatakse ka sünohulkade glossides ja näitelausetes olevaid sõnu.
Kolmandas sammus klassifitseeritakse kõik *wordneti* sünohulgad, kaasa arvatud need, mis eelmises sammus juba kindlaks määrati, kas positiivseks, negatiivseks või neutraalseks.
Teist sammu saab teostada erineva sünohulkade omavahelise kaugusega ja juhendatud masinõppemeetoditega. Kuna mitme klassifitseerija kasutamine osutus palju tulemuslikumaks, siis antud juhul kasutati nelja erinevat sünohulkade vahelist kaugust ja Rocchio ning tugivektormasina algoritme.
Neljandas sammus leitakse eelmises sammus loodud kaheksa klassifitseerija keskmine, millest saadaksegi lõplik tulemus.

1.3.2. Juhuslik liikumine

Juhuslik liikumine koosneb *wordneti* sünohulkade kui omavahel seotud punktide vaatlemises ning liikumisprotsessi iteratsioonide ehk mingite sammude jada korduva sooritamise käivitamisest. Iga iteratsiooni käigus võivad eelmises sammus saadud positiivsuse, negatiivsuse ja neutraalsuse skoorid muutuda.
Algoritmi idee seisneb selles, et mingi sõna kirjelduses olevad sõnad kannavad tõenäoliselt edasi kirjeldatud lekseemiga seotud meelestatust. Kuna definitsioonid võivad sisaldada ka erinevaid sünonüüme, tuleb seda teha sünohulga tasandil.
Kogu protsess viiakse läbi kaks korda ning tulemuseks saadakse kahe erineva skooriga (positiivsus ja negatiivsus) sünohulgad. Sellest hoolimata ei saa algoritmi tulemusena saadud numbrilisi väärtusi lõpliku skoorina kasutada, kuna need on selleks liiga väiksed ning näiteks

kõige positiivsemad sünohulgad on siiski suhteliselt neutraalse skooriga. Kuna sünohulgad on jaotunud niimoodi, et väga vähesed neist on väga kõrge positiivse ja negatiivse skooriga ning enamus pigem neutraalsed, tuleb neid väärtusi jaotuse järgi sobitada ning lõpuks muuta neid nii, et positiivsuse, negatiivsuse ja neutraalsuse summa võrduks alati ühega.

1.3.3. SentiWordNeti hindamine

Iga automaatselt koostatud andmestik vajab ka hindamist. Hinnangu andmiseks võrreldi käsitsi märgendatud sünohulki arvuti märgendatutega.
5 inimest valisid ja märgendasid kokku 1105 sünohulka. Sünohulkade 1-110 märgendamises osalesid kõik 5 inimest, et jõuda ühisele arusaamisele meelsuse märgendite suhtes. Seejärel märgendasid esimesed kolm sõltumatult sünohulgad 111-606 ning ülejäänud kaks märgendajat sünohulgad 607-1105.
Tulemuste kontrollimiseks järjestati eelnimetatud sünohulgad vastavalt positiivsuse ja negatiivsuse alusel. Seda tehti nii inimeste kui ka arvuti märgendatud sünohulkadega. Järjestusi võrreldi kasutades p-normaliseeritud Kendali T vahemaad. Tulemus on seda parem, mida lähemal on vahemaa nullile. Ehk siis kui p=0, on märgendused üks-ühele vastavuses, kui p=1, on märgendused üksteise suhtes vastupidised.
Tulemuseks oli positiivsete sünohulkade puhul 0,281 ja negatiivsete puhul 0,231. (Baccianella, Esuli, Sebastiani 2010)

1.4. SenticNet

Teine käesoleva magistritöö teoreetilises osas käsitletav *wordnet*'il põhinev meelestatuse analüüsi võimaldav ressurss on SenticNet[6]. Selle tutvustus käesolevas peatükis toetub Erik Cambria Robert Speer artiklile "SenticNet: A Publicly Available Semantic Resource for Opinion Mining".

SenticNet on avalikult kättesaadav ressurss, mis võimaldab meelestatust analüüsida kasutades tehisintellekti ja semantilisi võrke. SenticNet ei kasuta statistilisi masinõppe, vaid n.ö. argimõistusele omaseid (ingl *common sense*) vahendeid. SenticNeti loomisel olid abiks järgnevalt kirjeldatud mudelid.

[6] Vt http://sentic.net/ (18.05.2018)

1.4.1. ConceptNet

Inimesed toetuvad teineteisest arusaamiseks üksteisega suheldes suuresti jagatud taustateadmistele. Teadmine sellest, kuidas objektid on üksteisega seotud, inimeste eesmärgid nende igapäevaelus, sündmuste ja olukordade emotsionaalne tähendus - see kõik moodustab inimestele enesestmõistetavaks peetud info, mida nimetame argimõistuseks, asjad, mida inimesed peavad enesestmõistetavaks ja seega välja ei ütle (ingl *common sense*).

"The Open Mind Common Sense" projekt[7] on intuitsiooni võimaldamiseks tehisintellekti süsteemides ja rakendustes sellist infot vabatahtlikelt kogunud aastast 2000. ConceptNet[8] kujutab endast suunatud graafi, milles tipud on mõisted ning neid ühendavad kaared argimõistuslikud seosed nende vahel.

1.4.2. Tundeväli

Tundeväli ehk AffectiveSpace on n-mõõtmeline vektorruum, mis on kokku pandud ConceptNetist ning WordNet-Affectist[9]. Viimane nimetatuist on keeleteaduslik ressurss tunnetega seotud teadmiste leksikaalsel kujul esitamiseks.
Pärast vastavatest ressurssidest pärinevate lemmade kokkusobitamist tehakse saadud maatriksil singulaarsete väärtuste dekompositsioon ning visatakse välja andmestikust need elemendid, mis eriti palju ei varieeru. See annab tulemuseks kahemõõtmelise ruumi, mida nimetatakse tundeväljaks (ingl *affective space*), kus erinevad vektorid tähistavad erinevaid moodusi erinevate mõistete ja emotsioonide binaarseks eristamiseks. Seega kipuvad sarnast emotsiooni kandvad mõisted olema tundeväljas üksteisele üsna lähedal. Mõistete sarnasus ei olene mitte nende asukohast eelnimetatud ruumis, vaid vektorite omavahelisest nurgast. Näiteks "ilus päev", "sünnipäevapidu", "naerma", "kedagi õnnelikuks tegema" asuvad vektorruumis üksteisele üsna lähedal ning mõisted nagu "ennast halvasti tundma", "pisaraid valama" jne asuvad vektorruumi teises otsas.

1.4.3. Emotsioonide liivakell

Emotsioonide liivakell on mudel, mis põhineb ideel, et inimmõistus on eriressursside kogum ning et erinevad emotsionaalsed seisundid on osade ressursside sisse- ja teiste väljalülitamise tagajärg. Näiteks viha näib sisselülitavat just need osad, mis aitavad meil reageerida kiiremini

[7] https://en.wikipedia.org/wiki/Open_Mind_Common_Sense (18.052018)
[8] http://conceptnet.io/ (18.05.2018)
[9] http://wndomains.fbk.eu/wnaffect.html (18.05.208)

ja suurema jõuga, samal ajal mõistusega reageerimist võimaldavaid ressursse pärssides. Antud mudel on kasulik eelkõige emotsioonide tuvastamisel, arusaamisel ning väljendamisel suhtluses inimese ja arvuti vahel. Emotsioonide liivakella puhul ei ole tundeseisundid klassifitseeritud, vaid paigutatud nelja samaaegselt toimivasse ja üksteisest sõltumatusse mõõtmesse: meeldivus, tähelepanu, tundlikkus ning võimekus. Näiteks:

- meeldivus - kasutaja on pakutud teenusega rahul
- tähelepanu - kasutaja on pakutud informatsioonist huvitatud.
- tundlikkus - kasutaja tunneb end kasutajaliidest kasutades mugavalt.
- võimekus - kasutaja on süsteemi kasutamiseks piisavalt ettevalmistatud.

Kõiki neid nelja tunnetega seotud mõõdet iseloomustab 6 aktiveerumistaset, mis näitavad kogetud või tajutud emotsiooni tugevust. See teeb siis kokku 24 erinevat märgendit.

Tunnetega seotud mõõdikute koosesinemine võimaldab keerukamate emotsioonide seotud info loomist. Näiteks *armastus* on meeldivuse ja võimekuse positiivsete ning *pettumus* tähelepanu ja meeldivuse negatiivsete väärtuste summa.

1.4.4. SenticNet ja selle arendus

SenticNeti loomise eesmärgiks oli luua kogum sageli kasutatavatest mõistetest, millel on väga tugev positiivne või negatiivne polaarsus.
Seega kui näiteks SentiWordNet sisaldab ka nullpolaarsusega, ehk neutraalseid lekseeme, siis SenticNetist on need eemaldatud. Teiseks erinevuseks on see, et SenticNetis hoitakse meelestatuse info ühe ujukoma arvuna, mis on vahemikus -1.0-1.0. See teeb semantilise võrgu kuvamise palju lihtsamaks. Seega näiteks mõistetel “austust välja näitama” ja “hea tehing” on skoor üsna 1.0 lähedal ja “lahti lastud olema” ja “kontrolli kaotama” -1.0 lähedal.

1.4.5. Mõiste polaarsuse defineerimine

SenticNetis defineeritakse mõiste polaarsus emotsioonide liivakellast saadud märgendite summana. Kuid kui meeldivust ja võimekust saab väljendada nii positiivselt kui ka negatiivselt, on tähelepanu ja tundlikkus ainult ühesuunalise polaarsusega. Mõiste polaarsuse väljaarvutamine põhineb eeldusel, et tundeväljas olevate mõistete kaugused üksteisest on seotud nende polaarsuse erinevusega.

Iga liivakellas oleva tundega seotud mõõtme jaoks otsitakse mõisteid, mis on semantiliselt korrelatsioonis positiivsete väärtustega ning mittekorrelatsioonis negatiivsetega ja vastupidi.

Näiteks meeldivuse suhtes positiivsete mõistete leidmiseks otsime neid, mis on semantilises korrelatsioonis rõõmu ja rahuga ning samal ajal mittekorrelatsioonis kurbuse ja leinaga. Negatiivse polaarsusega mõistetega toimitakse täpselt vastupidi.

Selleks kasutatakse kahte erinevat võtet: segamine (ingl *blending*) ning spektraalne assotsatsioon. Esimeses võttes kasutatakse ära erinevatest allikatest pärinevate andmete kattumist. Sedasi saab omavahel kombineerida eri valdkondade, nagu näiteks meditsiini-, geoloogia- ja rahandusealaseid teadmisi. Seega annab omavahel kombineerida üldteadmisi ConceptNetist ning emotsioonidega seotud teadmisi Wordnet-Affectist.

1.4.6. Spektraalne assotsiatsioon

Spektraalne assotsatsioon koosneb erinevatele, n.ö. võtmetähtsusega mõistetele nagu hea või huvitav väärtuste omistamisest ning nende järgi erinevate seoste laialilevitamises. Seda võib näha kui alternatiivset meetodit mõistetele emotsionaalse väärtuse andmisest, mis ei sõltu välistest ressurssidest, nagu näiteks Wordnet-Affect.

Pärast eelkirjeldatud võtete kasutamist tuleb neid tervikliku ressursi huvides ümber korraldada. Näiteks tuleb võimalike vastuolude vältimiseks eemaldada duplikaatmõistetest need, millel on väiksem polaarsuserinevus. Suurema polaarsuserinevusega mõisted annavad enamasti suurema usaldusväärsuse.
Selleks, et SenticNet oleks arvuti abil kergesti loetav, viiakse see XML-formaati.

1.4.7. SenticNeti hindamine

Hindamiseks võrreldi SenticNeti SentiWordNetiga, kus kasutati 2000 patsiendi arvamust. Nende (ingliskeelsete) andmete põhjal andis SenticNet palju parema tulemuse täpsusega (precision) 79% *vs* 53%. Saagiseks oli SenticNetil 58% ning SentiWordNetil 46%. F-skooriks oli seega 67% *vs* 49%. (Cambria, Speer 2010)

Kuigi SenticNeti loojatel õnnestus võrreldes SentiWordNetiga oluliselt parem tulemus, tasub sealjuures arvestada, et mõne teise valdkonna tekstide peal võib tulemus olla hoopis teine. Peale selle ei õnnestunud viidatud artiklist leida täpset meetodit, mille alusel analüüsitavad tekstid valiti ja kuidas SentiWordNeti ja SenticNeti tulemusi omavahel võrreldi.
Lisaks võib enda loodud ressursi võrdlemisel teisega olla siiski teatud määral kallutatust.

1.5. Emotsioonidetektor ja valentsisõnastik

Meelestatuse analüüsiga on tegeletud ka Eestis. Eesti keele tekstide analüüsimiseks on Hille Pajupuu jt Eesti Keele Instituudist loonud valdkonnast ja teksti tüübist sõltumatu polaarsuse ehk emotsioonidetektori. Antud peatükk tuginebki Hille Pajupuu, Rene Altrovi ja Jaan Pajupuu artiklile "Identifying Polarity in Different Text Types" (2016).

Kuna tänapäeval väheneb järjest näost näkku suhtluse osakaal ning kirjaliku oma seevastu jällegi suureneb, oli üks nende eesmärkidest luua sõnastik, mis aitaks ennustada, kuidas loetav tekst lugejale mõjub. Teiseks oli mõte luua ressurss, mis aitaks inimestel tohutus infohulgas soovi korral valida, kas lugeda positiivseid või negatiivseid tekste. Kolmas eesmärk, ja nende autorite jaoks peamine, oli muuta sünteeskõne emotsionaalsemaks ja loomulikumaks.

Tekstist meelestatuse info kättesaamiseks katsetati nii leksikaalset kui ka masinõppepõhist lähenemist.

Masinõppealgoritmi treenimiseks ja leksikaalse lähenemise täpsuse hindamiseks loodi kõigepealt EKI-s tekstikorpus. Kuna leiti, et meelestatuse polaarsuse kindlakstegemiseks on optimaalse pikkusega just lõik, siis koondati sinna lõigud erinevate ajalehtede ja rubriikide (sport, arvamus, kodundus, elu, kommentaarid, krimi, kultuur ja maailm) tekstidest.
Kolmel inimesel paluti üksteisest sõltumatult hinnata, kas mingi lõik on positiivne, negatiivne, neutraalne või vastuoluline. Tulemusena saadi korpusesse 4086 meelestatuse infoga märgendatud lõiku.

1.5.1. Leksikoni loomine Emotsioonidetektori jaoks

Lisaks korpuse loomisele pöörati EKI-s tähelepanu ka leksikoni loomisele. Kuna suured leksikonid kipuvad tihti sisaldama palju müra ja seega võivad meelestatuse analüüsimisel anda ebatäpsemaid tulemusi, otsustati väikse, sagedasti kasutatavatest sõnadest koosneva sõnastiku kasuks. Kuna selliseid sõnu leidub erinevates tekstides, on sellise sõnastiku eeliseks hea kasutatavus erinevate valdkondade tekstide peal.

Nad lõid kõigepealt põhisõnastiku, kuhu koguti 3015 sagedasemat eesti keeles kasutatud sõna. Seejärel paluti neljal eesti keelt emakeelena kõnelejal teineteisest sõltumatult hinnata nende polaarsust kas positiivseks, negatiivseks, neutraalseks või vastuoluliseks.

Sedasi saadi polaarsussõnastikku 317 positiivset ja 322 negatiivset sõna, ülejäänud jäid neutraalseks. Pärast antonüümide ja tuletustega täiendamist saadi sõnastikku 617 positiivse ja 730 negatiivse märgendusega sõna ning seda suuremat sõnastikku nimetavad nad valentsisõnastikuks. Etteruttavalt tuleb nimetada, et siinses töös kasutati esimest – polaarsussõnastikku.

Kuna eesti keel on morfoloogiliselt rikas ning ühe sõna erinevad vormid kannavad endas sageli erinevat meelestatust (näiteks abi (positiivne) ja abita (negatiivne)), ei hakatud sõnavorme lemmatiseerima. Valentsisõnastiku lõplikuks suuruseks sai eri vormidega 38 628 tekstisõne. Sõnastikust on välja jäetud homonüümsed vormid nagu näiteks mees, tänavat jne.

Seejärel võrreldi meelestatuse märgenditega tekstilõikude korpust saadud valentsileksikoniga, et uurida, kui palju lõigus olevaid sõnu on sõnastikus ning liideti lõigus olevate sõnade meelestatuse märgendid kokku. Näiteks, kui lõigus oli 1 positiivse ja 4 negatiivse märgendusega sõna, määrati lõik negatiivseks. Kui positiivseid ja negatiivseid sõnu oli ühepalju, klassifitseeriti lõik vastuoluliseks. Kui ei leidunud ei positiivseid ega negatiivseid sõnu, liigitati lõik neutraalseks. Niimoodi oli võimalik omavahel võrrelda korpuses oleva tekstilõigu käsitsi märgendatud meelestatust ning leksikoni baasil automaatselt määratletud tekstilõigu meelestatust.

1.5.2. Masinõppel põhinev lähenemine

Masinõppe põhise lähenemise katsetamiseks kasutati EKI-s eelpool mainitud Naive Bayes'i ja tugivektormasinate (svm) meetodeid. Kuna vastuolulised lõigud vähendasid olulisel määral ennustuste täpsust, jäeti need kõrvale ning kasutati kolmesuunalist klassifitseerimist: positiivne, negatiivne ja neutraalne.

1.5.3. Tulemuste kontrollimine

Tulemuste õigsuse kontrollimiseks võeti igast ajakirjanduse rubriigist 100 lõiku.
Leksikaalse ja masinõppepõhise lähenemise hindamiseks võrreldi tulemusi inimeste määratud hinnangutega. Mõlema täpsus jäi 75% lähedale, aga erinevad lähenemised andsid erinevat tüüpi tekstide peal erinevaid tulemusi. Näiteks kultuuritekstide peal andis leksikaalne lähenemine tunduvalt parema tulemuse, seevastu kommentaaride puhul oli masinõppepõhise lähenemise puhul täpsus palju suurem. Viimast võib seletada sellega, et kommentaarid on tavaliselt üsna lühikesed, tihti ühelauselised, ning sisaldavad sageli slängi ja lühendeid. Selliseid sõnu aga sõnastikes sageli ei leidu.

Leksikaalne lähenemine andis kõige parema tulemuse sporditekste analüüsides ning masinõppe täpsus oli kõige suurem arvamuslugude peal. (Pajupuu jt; 2016)

Tausta kirjeldamise kokkuvõtteks võib osutada, et nii nagu inglise keele puhul, on ka eestikeelsete tekstide analüüsimisel omad eelised ja puudused nii masinõppe- kui ka leksikonipõhisel lähenemisel.

2. SentiWordNeti skooride lisamine Eesti Wordnetile

Eelnevalt kirjeldatud andmebaasidest, mida eeskujuks võtta, valiti selle töö jaoks välja SentiWordNet. Vaatamata eelmistes peatükkides käsitletud puudustele on SentiWordNet oma struktuurilt üsna lihtne ja üheselt mõistetav. Lisaks ei eelda see, erinevalt SenticNetist, teiste keeleressursside peale *wordneti* olemasolu. Lisaks on SentiWordNeti puhul tänu oma ühemõõtmelisusele (st positiivsus-negatiivsus skaalale) tõenäosus, et sealsed meelestatuse skoorid eesti keele jaoks sobivad, oluliselt suurem.

Nimetatud ressursis on meelestatuse info kirja pandud numbriliste näitajatena. Sellepärast kasutatakse käesolevas magistritöös vaheldumisi mõisteid märgend, skoor ja näitaja, aga nende tähendus käesoleva töö kontekstis jääb samaks.

Kuna *wordnetis* on igal sünohulgal unikaalne identifitseeriv number, mis on erinevate keelte jaoks loodud *wordnet*'tidel sama, siis on teoorias meelestatuse märgendite ületoomine SentiWordNetist üsna lihtne. Praktiliste sammude tegemiseks – et kokku panna Eesti Wordneti sünohulgad ja nendele vastavad inglise SentiWordNeti meelestatuse märgendid – tuli teha eeltööd, et andmebaasid oleksid ühildatavad. Järgnevalt kirjeldataksegi andmebaaside formaate ja nende teisendusi, mida siinse töö jaoks tehti.

2.1. Formaatide erinevused

Esimeseks probleemiks, millega töös kokku puututi, oli Eesti Wordneti ja SentiWordNeti andmebaaside formaatide erinevus. Kuigi *wordneti* struktuur ja tööpõhimõte on eri keelte puhul enam-vähem sarnane, siis võimalusi, kuidas neid kirja panna ja arvutile loetavaks teha, on erinevaid. Nimelt kasutas Eesti Wordnet praeguseks ajaks juba vananenud töövahendi Polaris import-eksport formaati[10], SentiWordNet on aga CSV-formaadis.

Esimene neist on üks andmebaasiformaatidest, milles säilitatakse sünohulkade semantilised suhted ning omavahelised seosed. CSV on aga laialt levinud tekstifailil põhinev tabeliformaat, mida saab lugeda nii erinevate tabelarvutustarkvaradega kui ka muude programmide ja skriptidega. Kuna antud töö eesmärgi jaoks ei ole olulised semantilised suhted, vaid hoopis sünohulgad, sinna kuuluvad sõnad ning nende kohta käiv meelestatuse info, siis on CSV-formaat uue ressursi jaoks sobivam.

[10] https://doi.org/10.15155/1-00-0000-0000-0000-0011DL

Teine oluline asi, mida silmas pidada, on, et Princetoni Wordnetist, mille struktuurile toetuvad ka teiste keelte *wordnet*'id, on mitu erinevat versiooni: 1.5, 1.6, 1.7, 1.7.1, 2.1, 2.1, 3.0 ja viimane 3.1[11]. Erinevad versioonid ei ühildu aga omavahel täielikult, mis tähendab, et sünohulga identifikaatorid ehk numbrid, mis üheselt määravad sünohulga ning mis on eri keelte *wordnet*'ides enamasti samad, on eri versiooniti erinevad.

Eesti Wordnet oli lõputöö alustamise ajal (2016.a. sügisel) versiooninumbriga 73 ning põhines Princeton Wordnetil versiooninumbriga 1.5. Inglise SentiWordNet oma meelestatuse märgenditega kasutas aga Princetoni WordNeti versiooni 3.0. Seepärast oli tarvis leida moodus, kuidas erinevate versioonide sünohulkade identifikaatorid omavahel kokku viia.

2.2. Andmete ühendamine

Andmete omavahel kokkuviimiseks sai programmeerimiskeeles Python kirjutatud vastav skript, mis asub Github/i repositooriumis aadressil https://github.com/gerthjaanimae/estwn-sentiment-analysis. Programmeerimiskeele valiku tingis peamiselt selle lihtsus, kirjutatud lähtekoodi hea loetavus ning vajadusel suure hulga lisateekide (lisafunktsionaalsuse) olemasolu.

Kirjutatud skript kasutab Neeme Kahuski loodud ühendusfaile ning EuroWordNeti[12] Pythoni moodulit. Nimetatud ühendusfailid on failid, milles on ühes veerus kirjas Wordnet 1.5 ning teises WordNet 3.0 sünohulga identifikaator. Iga sõnaliigi jaoks on eraldi failid[13].
EuroWordNeti moodul on loodud Polaris import-eksport formaadis oleva Wordneti parsimiseks ning töötlemiseks[14].

Loodud skripti tööpõhimõte seisneb järgmistes sammudes.

Kõigepealt kogutakse ingliskeelse SentiWordNeti failist kokku sõnaliikide kaupa kõik sünohulgad ning nende meelestatuse märgendid. Seejärel luuakse eelnimetatud ühendusfailide abil sõnastikud (antud juhul on tegemist Pythoni andmestruktuuridega, mitte keele mõistes sõnastikega), mille võtmeks on Princetoni Wordnet 1.5 sünohulga identifikaator ning väärtuseks versioon 3.0 sünohulga identifikaator.

[11] https://wordnet.princeton.edu/download/old-versions (18.05.2018)
[12] http://projects.illc.uva.nl/EuroWordNet/ (18.05.2018)
[13] http://www.talp.upc.edu/content/wordnet-mappings-automatically-generated-mappings-among-wordnet-versions (18.05.2018)
[14] https://gitlab.keeleressursid.ee/nemee/eurown (18.05.2018)

Seejärel hakatakse EuroWordNeti mooduli abil Eesti Wordnetti analüüsima.
Esimeseks sammuks on EstWN-i sünohulga identifikaatori leidmine. Kuna ühendusfailis oli number veidi teisel kujul, täpsemalt alati kaheksakohaline, siis tuli leitud identifikaatorit töödelda. Näiteks kui EstWN-is oleks see number olnud 5849, siis ühendusfailis oleks see 00005849. Seega tuli kontrollida, kas arv on kaheksakohaline ja kui mitte, siis lisada selle algusesse vajalik arv nulle.
Seejärel saadaksegi ühendusfaili kaudu SentiWordNeti failist vastava sünohulga meelestatuse näitajad.
Väljundfaili kirjutatakse sõnaliik, sünohulga number, meelestatuse skoorid, sõnad ja näitelause või selgitus.

Järgnevas tabelis 1 on näitena välja toodud viis rida loodud väljundfailist:

Tabel 1. Näide loodud ressursi failist

Sõnaliik	identifikaator	pos_skoor	neg_skoor	sõnad	näited
a	01047874	0.75	0	hea; õnnelik; rõõmus	juhuse, sündmuse, olukorra kohta
a	01013961	0	0	viimane; viimne; lõplik	püsiv, kehtima jääv, mitte enam muutuv, muudetav v. parandatav
a	01649720	0.625	0.25	nooruslik	noor(t)ele omane; noorena näiv v. tunduv
a	01014953	0	0	keskmine	(hrl. kolme üksuse v. indiviidi korral:) ajaliselt (v. vanuselt) vahepealne
a	01594146	0	0	keskmine; aritmeetiline keskmine	matemaatilist, statistilist keskmist osutav v. väljendav

Tulemus asub Github-i repositooriumis failis nimega “senti-estwn.csv” (https://github.com/gerthjaanimae/estwn-sentiment-analysis/blob/master/result-files/senti-estwn.csv).

Selleks, et oleks võimalik tekkinud inglis- ja eestikeelset andmebaasi võrrelda, valiti sünohulga identifikaatoriks Princetoni Wordnet 3.0 vastavad numbrid ehk siis need, mida ingliskeelne SentiWordNet kasutab.

Võib öelda, et kogu selle osa juures osutus kõige keerulisemaks ja ajamahukamaks just EuroWordNeti mooduli tundma õppimine ja arusaamine, kuidas selle abil oleks kõige

mõistlikum andmeid kätte saada. Valitud CSV-formaat, mida kasutab ka SentiWordNet, osutus siiski oma struktuurilt oodatust lihtsamaks ja eesmärgi jaoks sobivamaks.

2.3. Ühendamise tulemused

Erinevate ressursside ühendamise tulemusel saadi uus ressurss, milles on Eesti Wordnetist pärit 57 556 sünohulgal olemas meelestatuse skoorid. Inglise SentiWordNetis oli kokku 117 660 sünohulka. Sellel, et ühendamisel saadi vasted veidi alla pooltele sünohulkadele, võib olla peamiselt kaks erinevat põhjust. Esiteks ei olnud kõikide sünohulkade identifikaatorid kirjas eelnevalt kirjeldatud ühendusfailides.

Teiseks on inglise ja eesti keeles selliseid mõisteid, millele teises keeles üheseid vasteid ei leidu. Kolmandaks põhjuseks on see, et Eesti Wordnet on jätkuvalt täienev leksikon.

Saadud tulemused võib jagada paradigmaatiliselt:

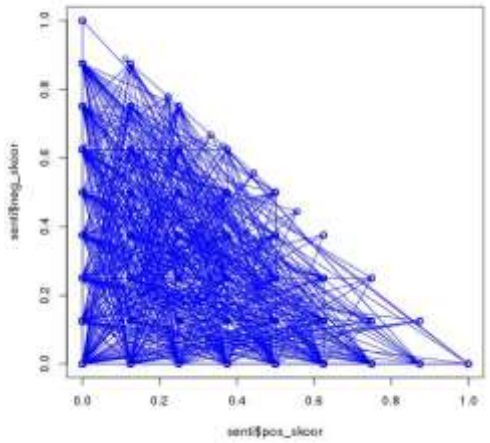
- tugevalt positiivseteks (1379),
- mõõdukalt positiivseteks (3067),
- neutraalseteks (46531),
- mõõdukalt negatiivseteks (2223),
- tugevalt negatiivseteks (1663),
- vastuolulisteks (2694).

Sellise tulemuse aluseks olid tugevalt positiivsetel ja negatiivsetel skoorid vastavalt suurem kui 0,25, mõõdukalt positiivsetel ja negatiivsetel skoorid 0,1 ja 0,25 vahel. Neutraalsetel oli objektiivsusskoor üle 0,9. Vastav skoor leiti valemiga:

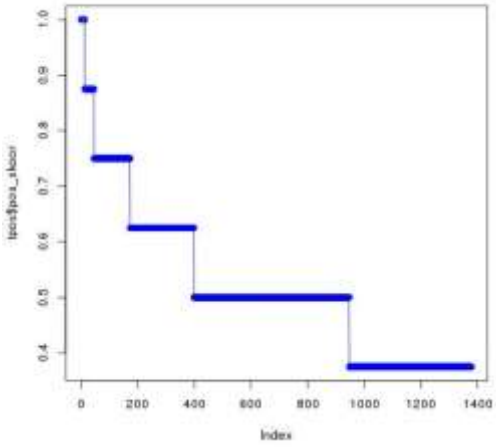
1 - positiivsus - negatiivsus ehk siis 1-st lahutati positiivsuse ja negatiivsuse näitaja.

Vastuolulistel olid nii positiivsus - kui ka negatiivsuskoorid üle 0,1.

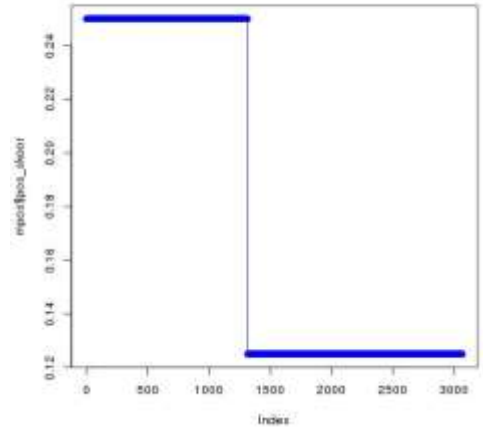
Selleks, et oleks hiljem andmeid lihtsam võrrelda, kirjutati statistikaprogrammis R skript, mis eelnevalt kirjeldatud alustel sünohulgad eraldi failidesse jaotab (nt tugevalt-positiivsed.csv, mõõdukalt-negatiivsed.csv jne). Lisaks joonistatakse vastava skripti abil eelnevalt kirjeldatud alustel ka diagrammid (vt allpool jooniseid 1–7) .



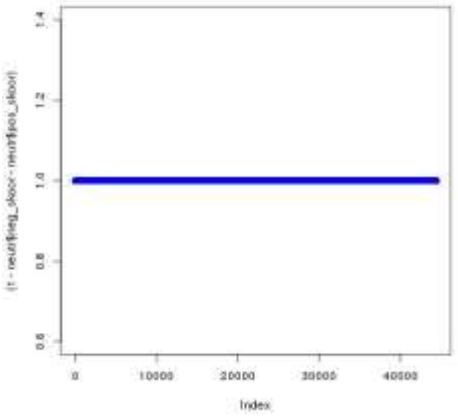
Joonis 1. Kõik meelestatuse skoorid



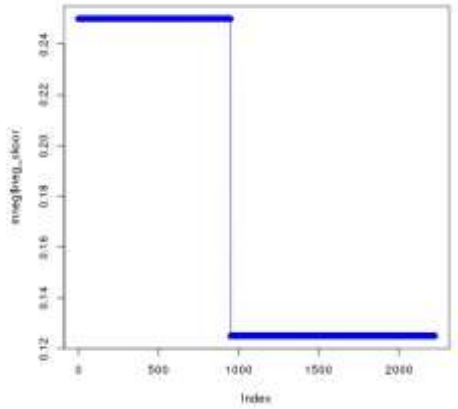
Joonis 2. Tugevalt positiivsed



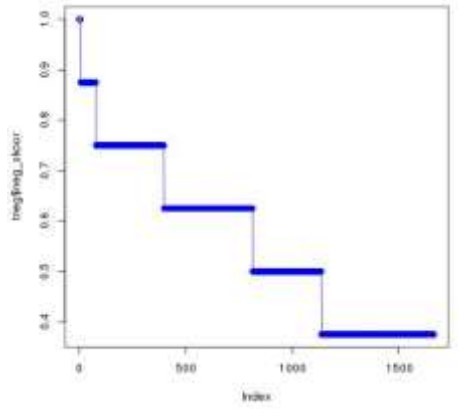
Joonis 3. Mõõdukalt positiivsed



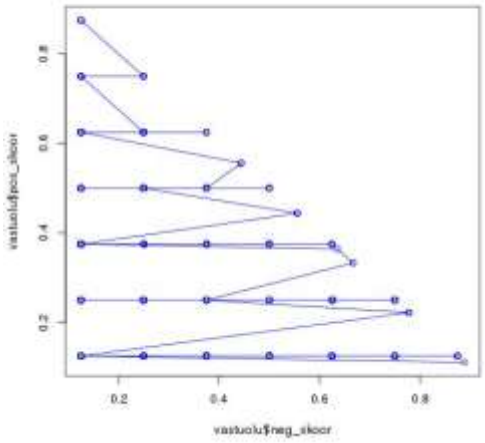
Joonis 4. Neutraalsed



Joonis 5. Mõõdukalt negatiivsed



Joonis 6. Tugevalt negatiivsed



Joonis 7. Vastuolulised

Nagu näha, on paradigmaatiliselt neutraalseid sünohulki kõige rohkem ning nagu illustreerib joonis 4, on kõigil neist neutraalsusskoorid võrdsed ühega, ehk siis võrduvad nende positiivsus- ja negatiivsusskoorid nulliga.
Kuna aga SentiWordNet lähtub põhimõttest, et info sõna neutraalsuse kohta on samuti oluline, on need andmed siiski olulised.

2.4. Võrdlus emotsioonidetektoriga

Kuna teine eesti keele jaoks väljatöötatud meelestatuse analüüsi võimaldav rakendus on peatükis 1.5 kirjeldatud Emotsioonidetektor, siis sai seal olevaid sõnu võrreldud loodud ressursiga, milles oli ühendatud EstWN ja SentiWordNet.

Kuna valentsisõnastik sisaldas ka palju eri sõnavorme, põhisõnastik (polaarsussõnastik) aga sõnalemmasid, siis sai võrdluse aluseks valitud sõnastikest viimane.
Selleks, et põhisõnastikku oleks skripti abil parem lugeda, teisendati see Exceli formaadist CSV-formaati.
Selleks, et põhisõnastiku ja meelestatuse skooridega Eesti Wordneti andmeid oleks mugav võrrelda, oli vaja need eelnevalt kokku panna. Sel otstarbel loodud Pythoni skript toimib järgmiselt:

- Kõigepealt loetakse sisse nii põhisõnastik kui ka SentiWordNeti põhjal loodud ressurss.

- Seejärel analüüsitakse sõna kaupa põhisõnastikku. Kui sõna magistritöö käigus loodud ressursis eksisteerib, siis kirjutatakse väljundfaili mõlemas ressursis olev meelestatuse info, kui mitte, siis kirjutatakse sõna standardväljundisse.

Et kindlaks teha, millised ja kui suur osa sünohulkadest ja põhisõnastikus olevaist sõnadest oma meelestatuse koha pealt vastuollu lähevad, sai Bashis kirjutatud skript, mis töötab järgmiselt:
Kõigepealt loetakse sisse eelnevalt loodud jaotuse põhjal loodud CSV failid (vt peatükk 2.3) näiteks tugevalt_positiivsed.csv. Seejärel võrreldakse ühendamise tulemusena lisatud EKI põhisõnastiku märgendeid. Kui seal leidub märgend, mida seal tõenäoliselt olla ei tohiks, ehk siis näiteks positiivsete puhul vastavalt neutraalne või negatiivne, siis lisatakse see vastuolude hulka.
Seejärel korratakse eelnevalt kirjeldatud samme kõikide teiste jaotuse tulemusena loodud failide peal.

Emotsioonidetektori põhisõnastikus oli kokku 3015 sõna. Eesti Wordneti 57 556 sünohulga meelestatuse märgenditega läksid omavahel vastuollu 906, st meelestatuse näitajad olid erinevad. Näiteks olid sõnad *salajane, varjatud* EstWN-is märgitud positiivseteks, EKI sõnastikus aga negatiivseks. Lisaks oli emotsioonidetektori põhisõnastikus 589 sellist sõna, mida uues loodud andmebaasis ei leidunud, näiteks *haisema ja halvenema*.

Põhjuseid, miks nii suur hulk sõnu Eesti Keele Instituudi sõnastiku ja EstWN-i vahel vastuollu läksid, on mitu:

1. “Ühe keele mõistete meelestatuse ülekandmine teisele keelele. Ühe keele- ja kultuuriruumis võib samal mõistel olla positiivne või negatiivne meelestatus, teises jällegi neutraalne või isegi vastupidine.” (Pajupuu, Altrov, Pajupuu 2016)
2. Paljud sõnad, mis EKI põhisõnastikus olid klassifitseeritud neutraalsetena, saaks EstWN-is liigitada mõõdukalt positiivsete või negatiivsete hulka. Seega on paljud neist tinglikult neutraalsuse ja positiivsuse ning negatiivsuse piiri peal. Täpsemalt on vastuolud eri meelestatuse kaupa kirjas tabelis 2. Sealt võib ka näha, et valdav enamus sõnadest on EstWN-is või EKI põhisõnastikus märgitud neutraalsetena.
3. Inimeste keeletaju on erinev, ehk siis üks inimene võib tajuda mingit sõna positiivsena, teine jällegi neutraalsena.
4. Üks ja sama sõna võib EstWN-is tähistada mitut erinevat mõistet. Näiteks oli sõna mõistmine põhisõnastikus klassifitseeritud neutraalsena. EstWN-is sisaldab seda sõna aga mitu sünohulka: ühes on see kui arusaamise sünonüümina, teises aga kui näiteks inimestevaheline mõistmine. Esimene oli märgendatud neutraalsena, teine

aga positiivsena. Kontrollimine, milline neist mõistetest mingil juhul sobib, on aga üsna keeruline.

Tabel 2. EstWN-i ja EKI põhisõnastiku meelestatuse vastuolud

Eesti Wordnetis	**EKI põhisõnastikus**	**vastuolude arv**
positiivsed	neutraalsed	155
positiivsed	negatiivsed	40
positiivsed	vastuolulised	0
neutraalsed	positiivsed	194
neutraalsed	negatiivsed	171
vastuolulised	positiivsed	68
vastuolulised	neutraalsed	59
vastuolulised	negatiivsed	69
negatiivsed	positiivsed	33
negatiivsed	neutraalsed	117
negatiivsed	vastuolulised	0

Täpsem info EstWN-i ja EKI põhisõnastiku vastuoludest on kirjas Githubis olevas failis "estwn-eki-contradictions.csv"[15].

2.5. Eesti Wordneti XML-faili ja SentiWordNeti ühendamine

Eesti Wordnet on pidevas muutumises nii oma mahu, kvaliteedi ja formaadi mõttes. Peale magistritöö praktilise osaga alustamist (2016.a. sügis) tuli EstWN-ist välja ka XML-formaadis versioon, mis vastab Princeton Wordnet versioonile 3.0, ehk siis sama, millega on seotud inglise SentiWordNet.

Kuna eelnevalt loodud ressursis õnnestus meelestatuse info lisamine EstWN-i sünohulkadele vaid osaliselt ja vaja oli saada meelestatuse skoorid ka ülejäänud sünohulkadele, siis prooviti SentiWordNetiga ühendada ka uus XML-formaadis olev EstWN.

[15] https://github.com/gerthjaanimae/estwn-sentiment-analysis/blob/master/result-files/estwn-eki-contradictions.csv

XML-formaadis olevas EstWN-is aga on sünohulkade identifikaatorid Princetoni Wordnetist märgitud erinevalt, vaatamata sellele, et versioon on sama, siis tuli kirjutada skript, mis nendele omavahelised vasted leiaks.
Selleks kirjutati programmeerimiskeeles Python skript, mis teeb HTTP-päringuid keeleressursside andmebaasidest.

Skripti tööpõhimõte on järgmine:

1) Kõigepealt kogutakse inglise SentiWordNeti failist kokku sünohulka identifitseerivad numbrid.
2) Siis eemaldatakse sünohulka identifitseeriva numbri algusest nullid ning lisatakse sõnaliik selle lõppu.
3) Seejärel tehakse kasutades saadud identifikaatorit iga sünohulga kohta HTTP-päringuid keeleressursside andmebaasist. Saadud päringute vastuseks saadakse JSON-formaadis andmed, millest saab kätte numbri, millele vastab sünohulk nimetatud XML-failis. Üldiselt on mõistlik HTTP-päringute vahele jätta veidi aega, et serverit mitte üle koormata, erinevate arvamuste kohaselt on selleks 5 kuni 30 sekundit. Kuna SentiWordNetis on sünohulki palju, töötab skript väga pikka aega (mitu päeva).
4) Saadud sünohulga number kirjutatakse koos SentiWordNetis oleva sünohulga numbriga faili. Põhimõtteliselt on tegemist eelnevalt kirjeldatud Neeme Kahuski ühendusfailide analoogse failiga.
5) Pärast seda võrreldakse saadud ühendusfaili ning esimeses etapis loodud meelestatuse märgenditega EstWN-i faili. Kogutakse kokku need sünohulgad, mille numbrit algselt loodud failis ei leidu.
6) Seejärel hakatakse EstWN-i XML-faili läbi vaatama. Kui sünohulka identifitseeriv number leidub eelmises sammus loodud failis, kirjutakse väljundfaili algne sünohulka identifitseeriv number, meelestatuse info ning EstWN-i sõnad ning näitelaused.

Kogu selle protsessi tulemuseks saadi täpselt sama struktuuriga fail nagu esimeses etapis (vt ptk 2.2.).
Töö käigus oli kõige ajamahukam just ootamine, kuna eelnevalt kirjeldatud põhjustel oleks mõistlik HTTP-päringute vahele mõistliku pikkusega paus jätta.
Kirjeldatud tulemusena suurenes eelnevalt tehtud ressurss 370 sünohulga võrra, sest ühenduslülisid sai väga vähe ning osad sünohulgad kattusid juba eelnevalt saadutega. Enamus uutest meelestatuse skooriga sünohulkadest olid omadussõnad.
Tulemus asub Githubi repositooriumis failis nimega “senti-estwn2.csv”[16].

[16] https://github.com/gerthjaanimae/estwn-sentiment-analysis/blob/master/result-files/senti-estwn2.csv

2.6. Ühendamise tulemuste reprodutseerimine

Ühendamisel saadud failid ning selle jaoks vajalikud skriptid asuvad Githubi repositooriumis https://github.com/gerthjaanimae/estwn-sentiment-analysis. Esimesed paiknevad kaustas result-files, teised aga kaustas *scripts*.

Tulemuste reprodutseerimiseks tuleb käivitada skript nimega *all-steps*.sh. Vastavas skriptis on ühtlasi ka kirjas, mis järjekorras ja kuidas loodud skripte käivitada tuleb. Lisaks on seal skriptide käivitamiseks vajalikud failid, mida automaatselt alla laadida ei õnnestunud. Need asuvad kaustas required-files.

Skriptide käivitamiseks peavad olema paigaldatud järgmised programmid ja lisateegid:

- Python3
- R (ka käsurealt skriptide käivitamiseks)

Lisateegid Python3 jaoks:

- requests
- BeautifulSoup
- LXML
- HTML-parser

Käesoleva magistritöö käigus kasutati Debian Linuxi distributsiooni, aga suure tõenäosusega töötavad need skriptid ükskõik millise Unixi-laadse operatsioonisüsteemiga.

3. Meelestatus ja sünonüümid

Igal sõnal võib olla palju stiilivarjundeid, mida isegi keelerääkija ei pruugi alati väga täpselt tajuda. Igas suhtlussituatsioonis tuleb otsustada, millist sõnalist valikut teha, et anda edasi mõeldud sisu ja vältida ebamugavusi, mis võivad tekkida valest sõnavalikust. Eriti keeruliseks muutub valik siis, kui inimese asemel tuleb otsus teha arvutil, näiteks masintõlkesüsteemil. Täielik ehk absoluutne sünonüümia, kui see üldse eksisteerib, esineb üsna harva. Absoluutsed sünonüümid peaksid olema igas kontekstis üksteisega asendatavad nii, et nende tõeväärtus, mõju suhtlemisel ning tähendus, kuidas see iganes ka defineeritud ei oleks, ei muutu. Osa filosoofe väidavad, et tõeline sünonüümia kui selline on võimatu, kuna seda pole võimalik defineerida ning tahtmatult jäetakse kõrvale sünonüümia teised vormid. (Edmonds, Hirst 2002)

Alain Cruse on ilmekalt öelnud selle nähtuse kohta ((1986) viidatud Edmonds, Hirst 2002 kaudu), et "loomulikud keeled jälestavad absoluutseid sünonüüme nii nagu loodus jälestab vaakumit, sest sõnade tähendused muutuvad aja jooksul pidevalt".

Parimal juhul kehtib absoluutne sünonüümia murretevahelises varieerumises ja tehnilistes terminites (Edmonds, Hirst 2002).

Seega on sõnad, mis on tähenduslikult üksteisele lähedal, pigem lähisünonüümid - väga sarnased, aga mitte identsed ning teatud määral varieeruvad otsese tähenduse, mõista andmise, meelestatuse ja rõhuasetuse poolest. Lähisünonüüme leidub eri keeltes palju ning neid leida ei ole üldse keeruline. Inglise keeles võib näiteks tuua sõnad *lie, falsehood, untruth, fib, ja misrepresentation*. Kõik need viitavad tõele mittevastavale infole, aga erinevad oma varjundi poolest. Sõna *lie* tähendab tahtlikku katset eksiteele viia, *untruth* valeinfo esitamist teadmatuse tõttu ning *fib* pigem hädavalet. (Edmonds, Hirst 2002)

Ka eesti keeles ei tule selliseid sõnu kaugelt otsida. Näiteks kirjeldavad *valetama, luiskama, võltsima, vassima* ja *susserdama* küll sama tegevust, aga üksteisest siiski veidi erineva nurga alt. Eesti keele seletava sõnaraamatu järgi tähendab sõna *luiskama* väljamõeldut rääkima, fantaseerima. Näitelaused: *Kas ta räägib tõtt või luiskab? Ära luiska, mis sa luiskad. Luiskab oma olematutest seiklustest.*

Vassima tähenduseks on EKSS-is aga: moonutama, võltsima, valesti esitama; puterdama, segi ajama; vigu tegema, eksima. Näiteks sündmusi, ajalugu vassima. *Vassitud andmed. Küll*

keerutab, valetab ja vassib. Vassis mis kole. Ajakirjanikud on nobedad vassima. Artiklis on ta mõtteid vassitud. Räägime otse, ilma vassimata. Ta on fakte meelega vassinud. Vassis vabanduseks midagi kokku.

Seega ei ole sünonüümide defineerimine ja üksteisest eristamine nii lihtne kui esmapilgul tunduda võib.

3.1 Pilootuuring

Kuna *wordnet* on mõistepõhine, mille moodustavad sünonüümsed sõnad, siis neile meelestatuse märgendite lisamise tulemusi kontrollides tekkis küsimus, kas ühte ja samasse sünohulka kuuluvad sõnad kannavad endas ikka sama meelestatust. Lisaks juba eelnevalt välja toodud valetamisega seotud sünonüümidele võiks lisaks nimetada sõnad *koer* ning *peni*, mis on küll sünonüümid, aga viimasel kipub olema pigem negatiivne varjund. Kui mõelda *wordnet*-tüüpi sõnastikust kui praktikas kasutatavast ressursist, siis erineva stiilinüansiga sõnad peaksid justkui olema erinevates sünohulkades. Sel juhul kaotaks aga mõistepõhisus sõnastikus oma algse mõtte.

Automaatselt Eesti Wordnetti meelestatuse skoore lisades oli oluline ka kontrollida, kas samasse sünohulka kuuluvad sõnad kannavad endas sarnast või erinevat meelestatust. Selleks ei piisanud aga ainult oma keelevaistust ning mõistlik oli kontrollida, mida keelekasutajad arvavad.

Selleks korraldati Google Formsi abil pilootuuring, mille käigus pöörduti keeleräkijate poole. Nendelt küsiti valitud sõnade kohta arvamust, millise meelestatusega võib nende jaoks mingi sõna olla.

Uuringu üleskutsele vastas 410 inimest. Vastajad olid vanusevahemikus 11- 80 aastat. Sõnad, mille kohta arvamust küsiti, olid järgmised: *kade, harrastuslik, amatöörlik, mahe, hüsteeriline, murelik* ja *kõmuline*. Küsitud sõnade arv oli üsna väike peamiselt sellepärast, et sedasi oli tõenäosus oluliselt suurem, et vastajad küsitluse ikkagi lõpuni täidaks.

Sõnade valik tehti järgmistel alustel:

- Esiteks samasse sünohulka kuuluvad sõnad, millel on potentsiaalselt üksteisest erinev meelestatus.

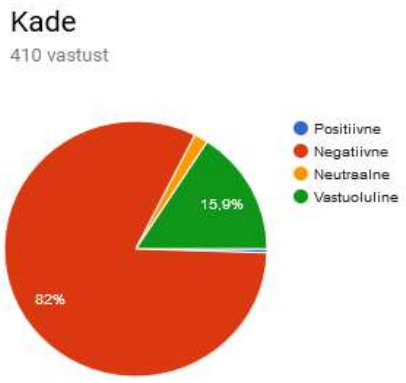
- Teiseks sõnad, mille meelestatust võivad keelerääkijad tajuda teisiti kui SentiWordNetil põhinevas ressursis.
- Selleks, et küsitlusele vastajad kasutaks vastamisel pigem oma keelevaistu, lisati küsitud sõnade hulka ka selline sõna, mille meelestatus peaks olema üldiselt üheselt mõistetav. Selleks sõnaks sai valitud *mahe*.

Küsitluse käigus küsitud sõnadest kuuluvad samasse sünohulka ainult kaks sõna - *harrastuslik* ja *amatöörlik*.

Järgnevates tabelites 3–9 on välja toodud iga küsitud sõna kohta kõigepealt SentiWordNetist saadud meelestatuse skoor ning seejärel inimeste arvamused nendes sisalduva meelestatuse kohta. Kui vastava sõna kohta on SentiWordNetis tähendusi mitu, siis on need mitmel real eraldi välja toodud. Samuti on viimases reas antud ka EKI hinnang, kui sõna kohta see oli olemas. Tabeleid illustreerivad neile vastavad diagrammid (vt jooniseid 8–14).

Tabel 3. Hinnang sõnale *kade.*

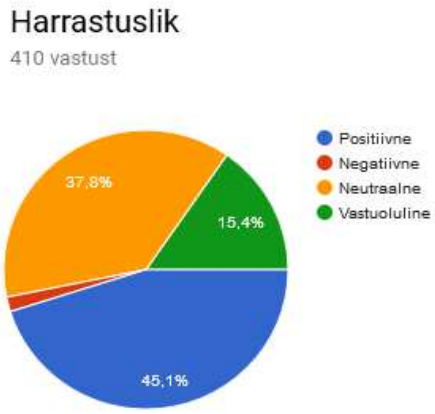
	Positiivne	Negatiivne	Neutraalne	Vastuoluline
SentiWordNet	0,5	0,375		
Küsitlus	2	336	7	65
EKI		negatiivne		



Joonis 8. Küsitluses antud hinnang sõnale *kade*.

Tabel 4. Hinnang sõnale *harrastuslik.*

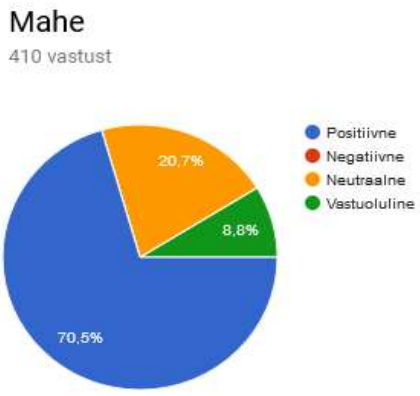
	Positiivne	Negatiivne	Neutraalne	Vastuoluline
SentiWordNet	0,375	0,125		
Küsitlus	185	7	155	63
EKI				



Joonis 9. Küsitluses antud hinnang sõnale *harrastuslik*.

Tabel 5. Hinnang sõnale *mahe*.

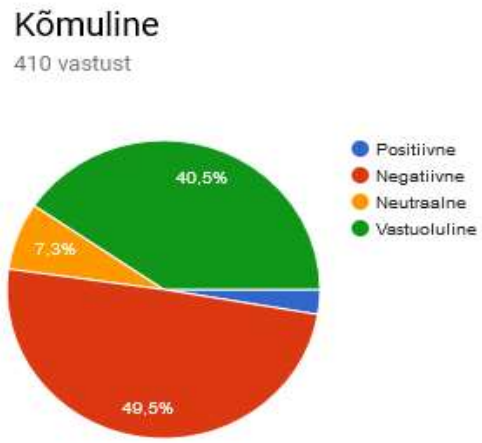
	Positiivne	Negatiivne	Neutraalne	Vastuoluline
SentiWordNet	leebe, sume 1 kõrvale meeldiv 0 vaevalt kuulda 0	0 0 0,125		
Küsitlus	289	0	85	36
EKI				



Joonis 10. Küsitluses antud hinnang sõnale *mahe*.

Tabel 6. Hinnang sõnale *kõmuline.*

	Positiivne	Negatiivne	Neutraalne	Vastuoluline
SentiWordNet	1	0		
Küsitlus	11	203	30	166
EKI				



Joonis 11. Küsitluses antud hinnang sõnale *kõmuline*.

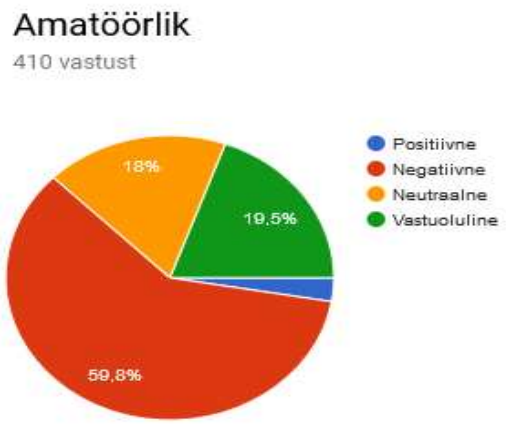
Tabel 7. Hinnang sõnale *murelik.*

	Positiivne	Negatiivne	Neutraalne	Vastuoluline
SentiWordNet	raske olukorra v tunnete kohta 0,25 muret tekitav 0 mure olev 0	0,5 0,75 0,875		
Küsitlus	11	252	71	76
EKI		negatiivne		

Joonis 12. Küsitluses antud hinnang sõnale *murelik*.

Tabel 8. Hinnang sõnale *amatöörlik.*

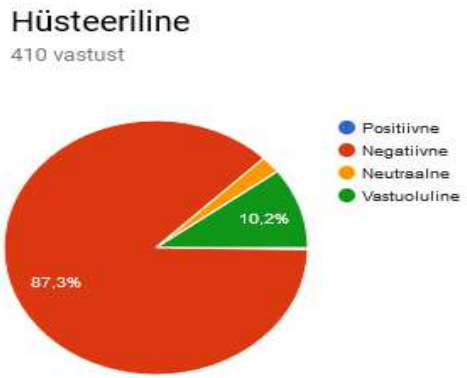
	Positiivne	Negatiivne	Neutraalne	Vastuoluline
SentiWordNet	0,375	0,125		
Küsitlus	11	245	74	80
EKI				



Joonis 13. Küsitluses antud hinnang sõnale *amatöörlik*.

Tabel 9. Hinnang sõnale *hüsteeriline.*

	Positiivne	Negatiivne	Neutraalne	Vastuoluline
SentiWordNet	0,375	0,375		
Küsitlus	1	358	9	42
EKI				



Joonis 14. Küsitluses antud hinnang sõnale *hüsteeriline*.

3.2. Küsitluse tulemuste tõlgendamine ja järeldused

Uuringu tulemus näitas, et samasse sünohulka kuuluvad sõnad võivad olla erineva meelestatusega. Näiteks sõnad *harrastuslik* ja *amatöörlik* kuuluvad samasse sünohulka ja on SentiWordNetis mõlemad positiivsema meelestatusega. Tabelitest 4 ja 8 ning neid illustreerivatest joonistest 9 ja 13 võib näha aga, et eesti keele rääkijad hindavad sõna *harrastuslik* positiivseks ja sõna *amatöörlik* valdavalt negatiivseks.

Siinkohal tuleks erinevusi *wordnet'*is olevate meelestatuse näitajate ja küsitlejate arvamuste vahel silmas pidades tähele panna, et kuigi näiteks *kõmuline* võib ingliskeelses kultuuris olla positiivsema tähendusega kui eesti keeles, arvestades ingliskeelset ütlust „There is no such thing as bad publicity" – „Ei ole olemas sellist asja nagu halb kajastus.", näivad käesoleva magistritöö koostaja keeletaju järgi *kade* ja *hüsteeriline* olevat siiski negatiivsed paljudes kultuuriruumides. Seega võiks oletada, et SentiWordNetis olevad erinevused keelekasutajate arvamustest ei ole ainult keelespetsiifilised ning probleem ei ole vaid ühe keele sõnade meelestatuse ülekandmine teisele keelele.

4. Arutelu

Käesolevas magistritöös rakendatud meetodil meelestatuse sõnastiku loomisel on kindlasti omad eelised, mis on omased leksikaalsele lähenemisele meelestatuse analüüsiks. Nendeks on peamiselt võimalus kasutada ressurssi ükskõik mis valdkonna tekstide peal ja selle kohene kättesaadavus, mis tähendab, et ei ole tarvis kulutada aega süsteemi treenimiseks, nii nagu tuleb teha näiteks masinõppepõhise lähenemise puhul. *Wordnet'*i kasutamine annab piisava leksikoni suuruse ja laiahaardelisuse ning tänu sellele, et sünohulgad on identifitseeritavad unikaalse identifikaatoriga, mis on eri keelte *wordnet*'ides samad, on võimalik ühe keele andmeid teise keelde üle tuua.

Samas esineb siinses töös kasutatud lähenemisel ka puudusi, mida oleks mõistlik arvestada. Esiteks seisneb üks probleem selles, et kuigi erinevate keelte *wordnet*'ides on sünohulgad ühendatud unikaalse identifikaatoriga ning seega ei ole tegemist mitte niivõrd sõnasõnalise tõlkimisega, vaid mõistelisega, siis leidub ikkagi paljusid mõisteid, millel on eri keeltes erinev meelestatus. Seda kinnitavad ka EKI valentsisõnastiku koostajad (Pajupuu jt 2016): "Näiteks kuuluvad eestlase jaoks positiivsete sõnade hulka leib, vaikne ja sõltumatu, seevastu hilinema ning vihmane on negatiivsed. Mõnes teises kultuuris võivad nimetatud sõnad olla neutraalsed või isegi vastandliku meelestatusega."

Teiseks on kõikide ressursis olevate sõnade meelestatuse näitajate kontrollimine väga aja- ja ressursimahukas. Lisaks on numbriliste skooride parandamise puhul probleemiks küsimus, mis alusel täpselt parandusi läbi viia.

Kolmas, lõputöö koostaja arvates olulisem ja *wordneti*-põhise lähenemise üldine probleem seisneb aga selles, et ühte sünohulka grupeeritud sünonüümid võivad olla erineva meelestatusega. Keeleteaduslikust kirjandusest on teada, et täissünonüüme eksisteerib suhteliselt harva (vt nt Cruse 1986; Apresjan 1992; Edmons, Hirst 2002). Käesoleva magistritöö käigus läbiviidud uuring kinnitas samuti seda, et ehkki sünohulga peaksid moodustama täissünonüümsed sõnad, siis võib neid pigem määratleda lähisünonüümseteks. Ehk siis üks sünonüümidest võib kanda endas neutraalset, teine aga negatiivset meelestatust või vastupidi.

Kuna *wordneti*-põhistes leksikonides on meelestatuse info sageli kirjeldatud skooridena, mitte lihtsalt märgenditena, siis tekib lisaks see probleem, et sünonüümid võivad olla küll sama meelestatusega, aga reaalsuses on nende kaalud erinevad - üks võib olla tugevalt, teine aga

mõõdukalt negatiivne. Näiteks kuuluvad samasse sünohulka sõnad nagu *joodik, alkohoolik, joomar, napsitaja, purjutaja, joomakaru, joomakõri, joomahaige, topsisõber, napsisõber, napsiarmastaja, lakkekauss, lakard, lakkekrants, pudelipaikaja, kõri, joomakoer, joomakrants* ja *parm*. Kuigi nimetatud sõnad tähistavad sama tüüpi inimest ning seega oleks nagu sünonüümid, on nendel siiski üsna erinevad varjundid. Olgugi et kõik need sõnad on intuitiivselt pigem negatiivse meelestatusega (kusjuures SentiWordNetil põhinevas ressursis olid need mingil põhjusel positiivse skooriga), siis ei ole *alkohoolik* ja *napsitaja* nii halvustavad kui *joodik*. *Lakkekrants* ja *parm* on jälle veelgi negatiivsema varjundiga.

Probleemi lahendamiseks aitaks see, kui lisaks sünohulgale saaks selle sees olevatele sõnadele eraldi meelestatuse näitajaid lisada. Teine võimalus oleks luua teistsuguse meelestatuse või varjundiga sõnadele uued sünohulgad. Tekib aga küsimus, kui palju ja mis alusel sünohulki üksteisest eraldada tuleks. Samuti kaotaks sellisel juhul *wordnet*'il põhinev lähenemine osaliselt oma mõtte, st tähendused ei saa olla liialt üle-eristatud.

Üks võimalus meelestatuse sõnastiku täpsemaks muutmiseks oleks ühendada Eesti Wordnet ja Emotsioonidetektori sõnastik, leides igale viimasest nimetatud sõnastikust pärit sõnale EstWN-i abil sünonüüme, antonüüme, hüperonüüme ja hüponüüme. Niimoodi saaks ju üsna suure leksikoni, mis arvestab ka keele spetsiifikaga. Kuna aga semantilised suhted ei pruugi tähendada sama meelestatust, on selle mõttekus töö koostaja arvates teatud määral küsitav.

Vaatamata sellele, et leksikaalse lähenemise üks suurtest eelistest on kasutatavus ükskõik mis valdkonna tekstide peal, on just sellega seoses üldiseks probleemiks, et üks sõna võib erinevates valdkondades või kontekstides olla erineva meelestatusega. Näiteks kui kirjeldatakse mõnda filmi või raamatut *huvitavana*, siis tähendab see enamasti positiivset. Kui aga mõne söögi kohta öeldakse *huvitav*, siis pahatihti on see aga negatiivse meelestatusega.

Lisaks on eestikeelsete tekstide puhul leksikaalse lähenemise kasutamine veidi problemaatilisem ka sellepärast, et eesti keelel on rikkam morfoloogia kui näiteks inglise keelel. Muidu positiivse varjundiga sõna võib abessiivi (ilmaütlevat käänet) kasutades omandada hoopis negatiivse meelestatuse. Seega tuleks lisaks eitustele ja eessõnadele teksti ka morfoloogiliselt analüüsida ning lisaks saadud sõnalemmadele uurida ka nende käändeid või muud vormilist infot.

Seega vaatamata teatud eelistele on *wordnet*'il põhinev lähenemine meelestatuse analüüsimiseks mitmes mõttes problemaatiline.

5. Edasised arendamisvõimalused

Magistritöö käigus loodud ressurssi on tulevikus võimalik mitmel erineval viisil edasi arendada.

Esiteks, kuna praegune versioon on ainult osaliselt märgendatud meelestatuse näitajatega, siis kindlasti oleks oluline lisada märgendus kogu Eesti Wordnetile. Selleks võiks juurde võtta ka teisi andmebaase, nt SenticNet, Wordnet-Affect jm. Kuna erinevates ressurssides on aga meelestatus märgitud erineval moel, siis võib nende omavaheline kombineerimine osutuda üsna keeruliseks.

Lisaks saaks Eesti Keele Instituudi Emotsioonidetektori jaoks loodud valentsisõnastikku laiendada, st kasutades ära wordnetis olevaid semantilisi suhteid nagu sünonüümia, antonüümia, hüperonüümia jne. Samuti võiks unistada, et Emotsioonidetektori kasutaja saaks mõlema andmebaasi meelestatuse näitajate kohta infot, juhul kui need omavahel vastuollu lähevad.

Peale selle on meelestatuse analüüsimiseks oluline kontekstitundlikkus. Näiteks sõnapaarid *õudselt ilus, jõle lahe, kohutavalt kaunis* on kahtlemata positiivsed, kuigi esimene sõnadest on eraldi vaadates negatiivne. Sellistel juhtudel võib isegi öelda, et negatiivse meelestatusega sõna suurendab positiivse sõna positiivsust. Selleks võiks näiteks mõnest tekstikorpusest otsida kõrvuti vastandliku meelestatusega sõnu ning moodustada nendest bigrammid. Nende teise sõna järgi saaks ehk esimesele ka antud kontekstis õige märgenduse anda. Välja selgitamiseks, kas ja kui hästi see toimib, oleks aga vaja üsna mahukaid korpuseuuringuid.

Võiks korraldada ka ulatuslikuma uuringu inimeste seas, et teada saada arvamusi nii-öelda keelerääkijalt. Kuigi oma eriala inimesed oleks kahtlemata pädevamad erinevate sõnade meelestatust hindama, oleks samuti hea, mida reaalsed keelekasutajad erinevatest sõnadest arvavad. Selleks saaks luua Google Formsi või mõne muu ankeetide jaoks mõeldud keskkonna abil mitu erinevat küsitlust ning küsitletu siis juhuslikult ühele neist suunata. Sedasi saaks piisavalt suure vastajate arvu korral tagada selle, et võimalikult paljud sõnad saaksid hinnangud ning samas ei oleks küsitletul tarvis ülemäära palju sõnu hinnata.

Kuigi magistritöö raames sai läbiviidud pilootuuring, mille käigus küsiti suvaliste keelekasutajate käest, millisena nemad mõnede sõnade meelestatust tajuvad, oleks hea, kui leksikoni vaataksid siiski läbi ka keeleteadlased või psühholoogiaga tegelevad inimesed. Kuigi

käsitsi läbivaatamine on väga aja- ja ressursimahukas, suurendaks see olulisel määral leksikaalse lähenemise puhul meelestatuse analüüsi täpsust. Sõnastiku käsitsi koostamise põhimõttele toetub ka Eesti Keele Instituudis loodud valentsisõnastik.

Saaks ka ära kasutada eesti keele seletavat sõnaraamatut, kus osadel sõnadel on juures märksõna vulgaarne või halvustav. Need saaks omakorda siduda EstWN-i semantiliste seostega.

Need on ainult mõned ideed, kuidas loodud keeleressurssi edasi arendada võiks.

6. Kokkuvõte

Magistritöö eesmärk oli luua keeleressurss, mis võimaldaks rakendada meelestatuse analüüsi Eesti Wordneti abil. Eesmärk sai täidetud. Eesti Wordneti sünohulkadele sai lisatud meelestatuse skoorid, mis õnnestus osaliselt - 86 000 sünohulgast said meelestatuse märgenduse 57 556 sünohulka. Selline tulemus näitab, et automaatsete keeletöötlusvahenditega on suhteliselt hõlpsasti teostatav täiesti uue keeleressursi loomine. Muidugi kaasnevad sellega teatud probleemid ja neid käsitleti eelnevates peatükkides.

Uue keeleressursi automaatne loomine tähendab väga ressursirohket ja mahukat tööd. Peale ühendamise enda oli vaja põhjalikult uurida ühendatavaid osi, kuidas need toimivad ja kuidas oleks neid kõige mõistlikum kasutada.

Lisaks uue keeleressursi loomisele viidi läbi pilootuuring selgitamaks, kas samasse sünohulka kuuluvad sõnad kannavad endas sama meelestatust. Uuringu tulemused kinnitasid, et alati ei pruugi see nii olla. Edasised keeleteaduslikud uuringud selles vallas oleksid vajalikud.

Kuigi käesolevas magistritöös kasutatud lähenemisel on nii mõnedki puudused, nagu eri keelte varieeruv meelestatus sama mõiste suhtes ning samasse sünohulka kuuluvate sõnade erinev meelestatuse määr, tasub ehk tulevikus antud teemaga siiski edasi tegeleda.

Kasutatud kirjandus

Apresjan, J. 1992. Systemic lexicography. — Euralex-92. Proceedings. Part I. Tampere, 3–16.

Baccianella, S.; Esuli, A.; Sebastiani F. 2010. SENTIWORDNET 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. http://nmis.isti.cnr.it/sebastiani/Publications/LREC10.pdf (18.05.2018)

Cambria, E.; Speer R.; Havasi, C.; Hussain, A. 2010. SenticNet: A Publicly Available Semantic Resource for Opinion Mining In: Commonsense Knowledge: Papers from the AAAI Fall Symposium, Menlo Park, CA, USA: AAAI Press. AAAI Fall Symposium, 11.11.2010 - 13.11.2010, Arlington, VA, USA, pp. 14-18. http://sentic.net/senticnet.pdf. (18.05.2018)

ConceptNeti koduleht: http://conceptnet.io/ (18.05.2018)

Cruse, A. 1986. Lexical Semantics. Cambridge University Press.

Edmonds, P.; Hirst, G. 2002. Near-Synonymy and Lexical Choice. Computational Linguistics. Volume 28 , No. 2. pp. 105-144.

EuroWordNet projekti koduleht: http://projects.illc.uva.nl/EuroWordNet/ (18.05.2018)

Orav, H.; Zupping, S.; Vare, K. 2014. Leksikosemantiliste suhete hägusus Eesti Wordnetis. Emakeele Seltsi Aastaraamat, 60 (2014), 171–194.

Princeton WordNet: https://wordnet.princeton.edu/ (18.05.2018)

Pajupuu, H.; Altrov, R.; Pajupuu, J. 2016. Identifying polarity in different text types. Folklore. Electronic Journal of Folklore, 64, 25−42. http://www.folklore.ee/folklore/vol64/polarity.pdf (18.05.2018)

Poola Wordnet http://plwordnet.pwr.wroc.pl/wordnet/ (18.05.2018)

Riikliku programmi „Eesti keeletehnoloogia 2011-2017“ https://www.keeletehnoloogia.ee/et/ekt-projektid/eesti-wordneti-taiendamine-2/eesti-wordneti-taiendamine-2 (18.05.2018)

SentiWordNeti koduleht: http://SentiWordNet.isti.cnr.it/ (18.05.2018)

SenticNeti koduleht: http://sentic.net/ (18.05.2018)

Vohra S. M.; Teraiya J. B. 2013. A Comparative Study of Sentiment Analysis Techniques. https://pdfs.semanticscholar.org/3f10/b006bab60c7f363bc03e72ad405d264b8d42.pdf (18.05.2018)

Word-Affecti koduleht: http://wndomains.fbk.eu/wnaffect.html (18.05.208)

Lisa

Magistritöö käigus loodud skriptid ja andmebaasid asuvad repositooriumis: https://github.com/gerthjaanimae/estwn-sentiment-analysis

Resume

Estonian Wordnet and sentiment analysis

The aim of this master thesis was to create a lexical resource for sentiment analysis in Estonian language and to find out how well does transfering sentiment tags from one language to another work. The choice was to use Wordnet, As it has been used quite successfully for this purpose in other languages, such as English and Polish and Estonian Wordnet today is quite considerable resource, containing over 86200 synsets. The main advantage of Wordnet is the fact that synonymous words of the same concept are grouped together into one synset. Thus by assigning the sentiment values to one synset it is possible to effectively tag multiple words at the same time.

The sentiment scores were derived from SentiWordNet 3.0 for English language. The result was partially successful. About 57000 synsets were tagged out of 86200.

As Emotsioonidetektor, compiled by Eesti Keele Instituut, is another tool for sentiment analysis for Estonian language, the words and their sentiment values were compared between each other.

As the other aim of the thesis was to find out if synonyms, words belonging to the same synset, do always have the same connotation, a pilot survey was carried out. Results of the survey confirmed that it is not always the case. Therefore it seems that the advantage mentioned before can be a disadvantage at the same time and Wordnet based approach has some issues to consider.

Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina Gerth Jaanimäe
(sünnikuupäev: 21.06.1989)

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose

Eesti Wordnet ja meelestatuse analüüs

mille juhendaja on Heili Orav,

1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.

2. olen teadlik, et punktis 1 nimetatud õigused jäävad alles ka autorile.

3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.

Tartus 21.05.2018