Tartu Ülikool

Humanitaarteaduste ja kunstide valdkond

Eesti ja üldkeeleteaduse instituut

Arvutilingvistika osakond

Karl Gustav Gailit

SPONTAANSUSE JA FORMAALSUSE KUI DIMENSIONAALSE TEKSTIMUDELI DIMENSIOONIDE AUTOMAATNE HINDAMINE VEEBITEKSTIDES

Bakalaureusetöö

Juhendajad Kadri Muischnek, Kristiina Vaik

Tartu 2021

Sisukord

1. Sissejuhatus

Tänapäevane maailm on suures osas läinud üle Internetti, mis on põhjustanud ka suure elektrooniliste tekstide tulva. Arvutilingvistikas ja keeleletehnoloogias on Interneti tekstid muutunud väga mahukaks ressursiks, kuna erinevate keelemudelite treenimiseks on vaja üha rohkem kvalitatiivseid andmeid. Kuid Interneti tekstidel kui ressursil on märgatud üht kitsaskohta: kui traditsioonilised tekstikorpused on loodud nii, et tekstide päritolu ja tekstiliigiline kuuluvus on juba korpuse loomise käigus teada, siis automaatselt veebist korjatud korpuste ehk veebikorpuste puhul on allikaid rohkem ning nende liigitamiseks puudub sobilik tekstiliikide taksonoomia.

Tekstiliikide klassifitseerimiseks on Vaik jt (2020) välja pakkunud dimensionaalse tekstimudeli, kus tekste liigitatakse dimensioonide ehk keeleliste tunnuste komplektide alusel mõõdetud tuumiknähtuste põhjal. Oma bakalaureusetöös käsitlen dimensionaalse tekstimudeli kahte dimensiooni, formaalsust ja spontaansust. Töö praktiline eesmärk on luua süsteem, mis automaatselt hindaks sisendiks antud tekstide formaalsust ja spontaansust. Praktiline eesmärk tingib ka kahte teoreetilist eesmärki. Esimene teoreetiline eesmärk on formaalsust ja spontaansust kirjeldavate tunnuste otsimine kirjandusest, teine eesmärk on nendest automaatsel hindamisel kasutatava komplekti koostamine.

Automaatse hindamise süsteemi loomiseks kasutan tunnuste komplekte ning eesti keele ühendkorpuse 2019. aasta veebitekstide alamkorpust (Ühendkorpus 2019). Automaatset hindamist valideerin kasutades testkorpust, mis sisaldab tekste inimeste formaalsuse ja spontaansuse hinnangutega. Inimeste hinnangutega võrdlemine annab teada, kuidas automaatne hindamine inimeste hinnanguga suhestub ning kuidas saaks automaatset hindamist edasi arendada.

Töö esimeses osas *Tekstiliikide klassifitseerimine ja selleks kasutatavad tunnused* tutvustan varasemaid Interneti tekstide liigitamise uurimusi, sealhulgas Vaigu jt (2020) dimensionaalset tekstimudelit ning selle kahte dimensiooni formaalsust ja spontaansust. Esimeses osas otsin ka varasemast kirjandusest formaalsust ja spontaansust kirjeldavaid tunnuseid ning valin leitud tunnustest komplekti, mida automaatse hindamise süsteemis kasutada. Töö teises osas *Meetod* loetlen valitud komplektide tunnused ning kirjeldan, kuidas hindasin nende esinemist tekstis. Töö kolmandas osas *Evalveerimine* võrdlen automaatse hindamissüsteemi hinnanguid inimeste antud

hinnangutega, analüüsin süsteemi hinnanguid ning pakun välja formaalsuse ja spontaansuse automaatse hindamise edasiarendamiseks järgmisi samme.

2. Tekstiliikide klassifitseerimine ja selleks kasutatavad tunnused

2.1. Tekstiliikide klassifitseerimine arvutilingvistikas

Veebitekstide klassifitseerimine on keeruline ülesanne, kuna veebitekstid on väga heterogeensed (Santini jt 2011; Sharoff 2021). Veebitekste uurides on näha, et tegu polegi ühe kindla varieeruva tekstiliigiga vaid hoopis mitme erineva liigiga, kasutustingimustest ajaleheartiklite ning jututubade vestlusteni.

Veebitekstide korpuste loomisel aga liike ei märgendata, kuna tekstiliigilist kuuluvust pole võimalik internetist korjates tuvastada. Korpustes on ainus teksti kirjeldav märgend teksti allika veebiaadress, mis ei võimalda jaotada tekste alamhulkadesse, mis sobiksid edasiseks lingvistiliseks uurimiseks. (Sharoff 2021)

Veebitekstide liigitamise keerukusel on mitmeid põhjuseid. Laippala jt (2021) ja Sharoffi (2021) järgi kipuvad varasemalt pakutud mudelid olema ebastabiilsed, kuna nad õpivad tekste liigitama suures osas tähemärkide *n*-grammide põhjal. Seega jääb peamiseks liigitamise aluseks sõnavara, mis põhjustab treeningkorpuse väliste tekstide palju kehvema ja ebatäpsema liigitamise. Lisaks ei kasuta mitmed eelnevad mudelid üldsegi leksikaal-grammatilisi tunnuseid, mis on aga teadaolevalt eri registrite vahel erinevad (Laippala jt 2021). Kuid peamine probleem veebitekstide puhul on asjaolu, et ei ole selget ülevaadet, millistesse tekstiliikidesse veebitekstid jagunevad. Ei olda kindlad nii tekstiliikide arvus ega ka liikide piirides. (Santini jt 2011; Laippala jt 2021)

Veebitekstide liigitamise probleemi lahendamiseks on välja pakutud mitmeid võimalusi. Sharoff (2021) on uurinud, kuidas suure veebikorpuse liigitamisel mõjutavad korpuse omapärad ja erinevad lingvistilised tunnused mudeli stabiilsust ehk võimet õigesti liigitada korpusevälised tekste. Laippala jt (2021) on tekstide registrite järgi liigitamiseks välja pakkunud masinõppel põhineva mudeli, mis kasutab sõnavara ja grammatilisi tunnuseid. Kasutati suurt veebikorpust ja

saadi üsna stabiilne, treeningmaterjalist sõltumatu mudel, mille keskmine F1-skoor oli 74,51% üle 26 registri. Vaik jt (2020) on veebitekstide liigitamiseks pakkunud lahenduseks välja dimensionaalse tekstimudeli, kus tekste jagatakse dimensioonide ehk keeleliste tunnuste komplektide alusel mõõdetud tuumiknähtuste kaudu. Pakutavas mudelis on dimensioone kaksteist: abstraktsus, afektiivsus, instrueerivus, informatsioonitihedus, spontaansus, formaalsus, impersonaalsus, ajalisuse olulisus, interaktiivsus, subjektiivsus, keerukus ja argumentatiivsus. Mudel on teoreetiline ning on võimalik, et seda edasi uurides ja empiiriliselt katsetades võib dimensioonide arv ja sisu muutuda. Kuigi dimensioonid on eraldiseisvad, võivad nende vahel tunnused osaliselt kattuda. Kui ühe dimensiooni kõik tunnused on ka teise dimensiooni tunnusteks, siis pole enam tegu iseseisva dimensiooniga. Oma töös proovin automaatselt hinnata dimensionaalse tekstimudeli kahte dimensiooni: spontaansust ja formaalsust.

Spontaansus väljendub kõige enam suulises kõnes, esineb see ka tekstides, peamiselt kontekstides, kus teksti kirjutamisel on olnud reaalajalisi piiranguid, näiteks jututubades, sõnumivahetuses või ka foorumipostitustes. Vaik jt (2020) oletavad, et nagu kõnes, väljendub kirjalikes tekstides spontaansus vigaderohkuse ja lihtsusena. Vigaderohkust on näha näiteks kirja-, trüki- ja keelevigades, nagu kirjavahemärkide ja sõnade vahel tühikute puudumine või kirjakeele normingutele mittesobiva käänd- või pöördtüübi kasutamine (näiteks *pere* vs. *peret*). Teksti lihtsus väljendub peamiselt lausestruktuuris, kuna kasutatakse, kas lühikesi lihtlauseid või lihtsa struktuuriga liitlauseid, enamasti rindlauseid aga ka lihtsaid põimlauseid. Lisaks leidub rohkelt lühendamist ja asendamist (näiteks *ks* vs. *x* ja *õ* vs. *6*).

Formaalsus seevastu kirjeldab teksti keelekasutuse ametlikkust. Vaik jt (2020) oletavad, et seda väljendatakse peamiselt leksikaalselt, kasutades viisakusväljendeid, nagu pöördumisi ja teietamist, ning vältides kõnekeelsust. Lisaks on formaalsete tekstide laused keeruka ehitusega ning kasutatakse palju nominalisatsioone.

2.2. Formaalsuse ja spontaansuse keelelised tunnused

2.2.1. Tunnuste valimise põhimõtte

Formaalsuse ja spontaansuse automaatseks hindamiseks otsisin esmalt tunnuseid, mille alusel dimensiooni esinemist tekstis hinnata. Otsisin tunnuseid peamiselt eestikeelsete tekstide

hindamiseks, kuid vaatasin tekste ka inglise keele kohta, kuna tunnused võivad keelte vahel kattuda.

Formaalsuse puhul vaatasin formaalseid tekste üldiselt kirjeldavaid artikleid ning tunnuseid valisin kahest artiklist. Esimene on eelnevalt tehtud uurimuste juures mainitud Sheikha ja Inkpeni (2012) artikkel ingliskeelsete tekstide formaalsuse ja mitteformaalsuse alusel klassifitseerimisest. Teine on Kerge ja Pajupuu (2010) artikkel tekstiliikidest kõnetehnoloogias ja keeleõppes, kus keskendutakse tekstide formaalsusele.

Spontaansuse puhul vaatasin spontaanset või mittespontaanset kirjalikku keelt esindavaid tekstiliike kirjeldavaid artikleid. Väga spontaanne tekstiliik on uue meedia tekstid, nagu jututubade ja foorumite tekstid (Muischnek jt 2011), ning mittespontaanne tekstiliik on lepingute tekstid (Reinsalu 2011). Otsisin tunnuseid ka suulist kõne kirjeldavatest tekstidest, vaadates Lindströmi ja Toometi (2000) artiklit suulistest narratiividest, kus võrreldakse spontaanseid suuliseid narratiive mittespontaansete kirjalike narratiividega nagu ilukirjandustekstidega. Sellest artiklist leidsin samuti mõningaid tunnuseid, mis sobivad ka tekstide formaalsuse kirjeldamiseks.

Artiklid ei käsitle spontaansust ja formaalsust täpselt nii, nagu seda käsitleb dimensionaalne tekstimudel. Formaalsust käsitlevad artiklid hõlmavad tihti formaalsuse alla ka keerukuse, impersonaalsuse ja abstraktsuse, mis on dimensionaalses tekstimudelis eraldi. Ka spontaansuse puhul oli leida tunnuseid, mis sobiksid pigem ebakeerukust või mitteformaalsust kirjeldama. Dimensiooni kirjeldavate tunnuste juurde selliseid tunnuseid, mis pigem kirjeldavad mõnda teist tunnust, ei lisanud.

2.2.2. Ühised tunnused

Ühisteks tunnusteks pean tunnuseid, mida saab kasutada nii formaalsuse kui ka spontaansuse hindamiseks. Enamasti, kui tunnus mõjutab formaalsust positiivselt, mõjutab see spontaansust negatiivselt. Sama kehtib ka vastupidi, kui tunnus mõjutab spontaansust positiivselt, mõjutab see formaalsust negatiivselt.

Sõnavara mitmekesisus

Sõnavara mitmekesisus (inglise keeles *Type-Token Ratio* ehk *TTR*) väljendab sõnavara varieerumist tekstis. See saadakse jagades unikaalsete sõnade arvu kõikide sõnade arvuga. Sõnavara mitmekesisust on kasutatud tekstide formaalsuse määramiseks, kus kõrge mitmekesisus näitab formaalsust ja madal mitteformaalsust (Sheikha, Inkpen 2012).

Spontaansuse osas olen teinud intuitiivse eelduse, kuna ei ole leidnud varasematest uurimustest ei kinnitust ega ka vastuväiteid. Eeldan, et sõnavara mitmekesisus mõjutab tekstide spontaansust vastupidiselt formaalsusele, ehk kõrge mitmekesisus näitab mittespontaansust ja madal spontaansust.

Keskmine sõnapikkus

Keskmine sõnapikkus tekstis on tunnus, mida Sheikha ja Inkpen (2012) kasutavad ingliskeelsete tekstide formaalsuse määramisel ning mis nende mudeli analüüsil osutus väga määravaks formaalsuse tunnuseks. Eesti keeles tuleb aga eristada pikki lihtsõnu ja liitsõnu, mis koosnevad lihtsatest ja lühikestest osasõnadest, kuid on tähtede arvu poolest pikad. Seega eesti keele puhul vaatan mitte keskmise sõna pikkust, vaid keskmise lihtsõna või liitsõna osa pikkust.

Ka selle tunnuse puhul olen teinud spontaansuse osas intuitiivse eelduse, kuna ei ole leidnud varasematest uurimustest ei kinnitust ega ka vastuväiteid. Eeldan, et keskmine sõnapikkus mõjutab tekstide spontaansust vastupidiselt formaalsusele, ehk mida lühemad on sõnad, seda spontaansem on tekst ja vastupidi, mida pikemad sõnad, seda mittespontaansem tekst.

Emotikonid

Emotikonid on väga levinud tunnus uue meedia tekstides, eriti jututubades, et väljendada lühidalt ja efektiivselt kirjutaja emotsiooni (Muischnek jt 2011). Mitteformaalsust väljendavad emotikonid, kuna need väljendavad subjektiivsust ja tundeid (Sheikha, Inkpen 2012). Netitekstides leidub mitmeid erinevaid emotikone: tähtedest ja kirjavahemärkidest moodustatuid (:) , :-P) (Wikipedia); erisümbolitest moodustatuid (¯_(ツ)_/¯ , (͡° ͜ʖ ͡°)) (Wikipedia, Looks.wtf); Unicode'i standardis (The Unicode Consortium 2021) olevaid ühe sümboli pikkuseid emojisid (😊, 🍺); foorumispetsiifilisi emotikone, mille puhul pannes koolonite vahele kindla nime, ilmub foorumis teksti asemel pildike (näiteks :maasikas:). Kõiki emotikonide tüüpe käsitlen ühe tunnusena.

Sõnavara

Nii formaalsel kui ka spontaansel kirjakeelel on oma sõnavara, mis on dimensiooni skoori hinnates tunnuseks. Mõlemal dimensioonil on sõnavara, mis mõjutab dimensiooni skoori nii positiivselt kui ka negatiivselt. Formaalne sõnavara on näiteks mitmesugused viisakusfraasid (näiteks *Lugupeetud härra või proua*) aga ka formaalsed asesõnad nagu teietamine (Sheikha, Inkpen 2012), mitteformaalne sõnavara on näiteks familiaarne (sealhulgas sinatamine), kõnekeelne või vulgaarne. Spontaansele sõnavarale on samuti omane kõnekeelsus, kuid spontaansust väljendavad ka mitmed lühemad vormid (nagu *külmik* ja *teler*).

2.2.3. Ainult formaalsuse tunnused

Ainult formaalsust väljendavad tunnused väljendavad ainult formaalsust või mitteformaalsust ning ei väljenda ei spontaansust ega mittespontaansust.

Nimisõnade ja nende laiendite kõrge arv

Mida rohkem on tekstis nimisõnu ja nende laiendeid, seda kõrgem on teksti formaalsus. Nimisõnade laiendite all on mõeldud omadussõnu, määratlejaid ja kaassõnu, sealhulgas mitmeid verbivorme nagu *hoolimata*. Verbid, määrsõnad, asesõnad ja asemäärsõnad aga vähendavad teksti formaalsust ning suurendavad abstraktsust ja mitmetimõistetavust. (Kerge, Pajupuu 2010)

Subjektiivsus, tunnete avaldamine

Subjektiivsus ja isiklike tunnete avaldamine väljendab tekstis mitteformaalsust (Sheikha, Inkpen 2012). Seega on mitteformaalsed meelestatud sõnad nagu omadussõnad *igav* ja *kaunis*, tajuverbid nagu *nägema* ja isiklikku arvamust väljendavad verbid nagu *arvama*.

Ühendverbid

Partikliga täiendatud verbid väljendavad mitteformaalsust (Sheikha, Inkpen 2012). Eesti keeles on ainult osa ühendverbe mitteformaalsed, nimelt sellised, kus partikkel on tähenduse poolest üleliigne, näiteks *ära kaotama*.

Esimene, teine ja kolmas isik

Kolmas isik väljendab tekstis formaalsust nii asesõnana kui ka verbi isikuna. Esimene ja teine isik väljendavad mitteformaalsust, välja arvatud juhul, kui kasutatakse formaalseid asesõnu. (Sheikha, Inkpen 2012) Eesti keeles on teise isiku ainsuse formaalseks asesõnaks *Teie* ning ka verbidel kasutatakse teise isiku mitmust.

Impersonaal ja teised tegevussubjekti varjamise või tahaplaanile jätmise keelelised vahendid

Impersonaal ja teised tegevussubjekti varjamise või tahaplaanile jätmise keelelised vahendid väljendavad keeles formaalsust ning personaal ja teised tegevusubjekti väljatoomise vahendid väljendavad mitteformaalsust (Sheikha, Inkpen 2012). Eesti keeles on kõige tavalisem tegevussubjekti varjamise või tahaplaanile viimise vahend impersonaali ehk umbisikulise tegumoe kasutamine.

Kindla kõneviisi enneminevik ning *nud*-partitsiip

Mitteformaalsetes tekstides kasutatakse vahendatuse väljendamiseks kindla kõneviisi enneminevikku ja *nud*-partitsiipi (Lindström, Toomet 2000). *nud*-partitsiip väljendub ainult morfoloogilise tunnusena, kuid kindla kõneviisi enneminevik, ja kõik teised liitajad, on nii morfoloogiline kui ka süntaktiline tunnus.

vat-tunnuseline kaudne kõneviis

Suulises kõnes kasutatakse *vat*-tunnuselist kaudset kõneviisi ainult formaalses suhtlussituatsioonis (Lindström, Toomet 2000). Kõnekeeles on kaudset kõneviisi võimalik väljendada ka da-infinitiiviga.

2.2.4. Ainult spontaansuse tunnused

Ainult spontaansust väljendavad tunnused väljendavad ainult kas spontaansust või mittespontaansust ning ei väljenda ei formaalsust ega mitteformaalsust.

1. isiku asesõnad

Esimese isiku asesõnade sage kasutus on omane spontaansele suulisele narratiivile, kuna kõneleja viitab endale suhteliselt sagedaselt kasutatud uuritavas materjalis. Peamiselt on tegu esimese isiku

ainsuse asesõnaga *mina*, kuna mitmuslikke pronoomeneid kasutatakse vähem kui ainsuslikke. (Lindström, Toomet 2000)

Taju- ja tunnetusverbid

Suulistes narratiivides on kasutatud kirjalike narratiividega võrreldes ohtralt taju- ja tunnetusverbe nagu *vaatama*, *mõtlema* ja *tahtma* (Lindström, Toomet 2000).

Partiklid

Jututubade tekstides on sagedased partiklid ehk lühikesed sõnad, mis on sageli lühenenud sõnavormid ja mis on kas terviklikud lausungid või lausungi osana üldlaiendid. Partikleid on viite tüüpi. Esiteks dialoogipartiklid nagu *ok* ja *aa*, mis väljendavad, et sõnum on vastu võetud või ollakse kuuldel. Teiseks afektiivsed partiklid nagu *ups* ja *irw*, mis väljendavad teksti autori subjektiivset arvamust, meeleolu ning tundeid. Kolmandaks aktiivsed suhtluspartiklid, mille eesmärk on püüda tähelepanu, näiteks *kle* ja *tsau*. Neljandaks piiripartiklid, mis seostavad algava tekstiüksuse ülejäänud tekstiga, näiteks küsipartikkel *ve*. Viiendaks ja viimaseks on toimetamispartiklid nagu *hmm*, mis näitavad, nagu suulises kõnes, et teksti kirjutaja mõte on takerdunud või on toimunud mõtteis pööre. (Muischnek jt 2011)

Toorlaenud

Spontaansetes internetitekstides leidub mitmeid toorlaene. Neid leidub rohkem erialatekstides, nagu uudisgruppides ja foorumites, ning eesti keeles on enamasti tegu inglise keelest võetud erialasõnavaraga, kuid leidub palju ka vene ning teiste võõrkeelte laene. Enamasti jäävad laenud ühe sõnavormi või fraasi piiridesse, kuid leidub ka terveid võõrkeelseid lausungeid muidu eestikeelses tekstis. Sõnavormid võivad olla nii keelenõuetele vastavalt vormistatud tsitaatsõnad, võõrapärase kirjapildi kuid eestikeelse käändelõpuga toorlaenud või eesti keelde mugandatud kirjapildiga sõnad. Leidub ka võõrkeelseid partikleid, eriti inglise päritolu sõnu. (Muischnek jt 2011)

Kirjavahemärkide kasutamata jätmine

Jututubade tekstides on levinud kirjavahemärkide kasutamata jätmine. Jäetakse ära nii komasid kui ka lauselõpumärke, peamiselt punkti. (Muischnek jt 2011)

Sümboliasendused

Sümboliasendused on taotluslikud muudatused sõnas, kus üksiku tähe või tähekombinatsiooni asemel kasutatakse teisi tähti, numbreid või sümboleid. Levinud on täpitähtede asemel numbrite kasutamine, näiteks *õ* asemel *6*, mis võib olla tingitud kas puudu olevatest klahvidest või lihtsalt soovist tähti sõnas asendada. (Muischnek jt 2011) Seda tunnust leidub rohkem vanemates internetitekstides ning uuemates vähem.

Kirjavead

Kirjavead on spontaanses tekstis väga levinud, eriti jututubades ja teistes suhtluskeskkondades. Selle peamine põhjus on, et tekst kirjutatakse ja saadetakse kiiresti ja ilma läbi lugemata ning parandamata. Levinud on vead sõna tähtedes, näiteks *atpeek* sõna *apteek* asemel, aga ka kirjavahemärkide ning emotikonide kokkukleepumine ülejäänud tekstiga, näiteks koma järel tühiku puudumine. (Muischnek jt 2011)

Kordused

Uue meedia tekstides korratakse sageli tähti, silpe ning isegi terveid sõnu, et rõhutada nende sisu (Muischnek jt 2011). Üksiku tähe kordus on näiteks *eiiiiiiiii* ning sõnakordus on *väga väga VÄGA*. Samuti korratakse lauselõpumärke. Kui küsimärgi ja hüüumärgi kordamine intensiivistab neid kirjavahemärke, siis punktide kordamine täidab mõttepunktide rolli, ehk kui punkte on kolm või enam, täidavad need alati mõttepunktide rolli ning lause lõpu funktsiooni ei intensiivistata.

Kirjakeelele mittevastavad tähesuurused

Internetis on suhtluskeskkondades, eriti sotsiaalmeedias ja jututubades, levinud normeeritud kirjakeele suur- ja väiketähe reeglite mitte järgimine. Lauseid ja pärisnimesid alustatakse väikeste tähtedega ning suurtähti kasutatakse läbivalt, et sõnu rõhutada. (Muischnek jt 2011)

Mineviku liitajad

Suulistes narratiivides on täheldatud suuremat mineviku liitaegade kasutamist võrreldes ilukirjandusega (Lindström, Toomet 2000).

Tarindid

Mittespontaansete tekstide üks esinemisvorm on lepingutekstid, mida on uurinud Reinsalu (2011). Tema väitel on lepingutekstid sisutihedate ja keerukate ülesehitustega lihtlausetega. Nende keerukus on tingitud pealausetesse lisatud tarinditest, mille hulka kuuluvad infiniit- ja partitsiiptarindid, nominalisatsioonid, adverbalisatsioonid ja predikaadita tarindid.

Propositsioonid

Reinsalu (2011) uuritud lepingutekstid kui mittespontaansete tekstide üks esinemisvorm on kõrge süvastruktuuri elementaarlausete, mis on moodustatud süvastruktuuri predikaadist ja selle argumentidest, ehk propositsioonide arvuga. Seega mittespontaansed tekstid on väga sisutihedad.

Pikad lihtlaused

Mittespontaansete tekstide ühe esinemisvormi lepingutekstide rohked propositsioonid ja tarindid tähendavad, et tekstiliigi laused on pikad. Pindstruktuuri poolest on tegu lihtlausetega. (Reinsalu 2011) Reinsalu näidetest võib aru saada, et tegu on lihtlausetega, kus on rohkelt sisestatud struktuure, seega tarindite tõttu väga pikad lihtlaused on mittespontaansuse tunnuseks.

2.2.5. Tunnuste alamhulga valimine hindamiseks

Dimensioonide hindamiseks valisin leitud tunnustest alamhulga, mida kasutada dimensioonide hindamise katse realiseerimiseks. Esmalt kontrollisin üle, et ükski tunnus ei kattuks mõne dimensionaalse tekstimudeli dimensiooniga. Täheldasin, et subjektiivsus ja tunnete avaldamine ei sobi formaalsuse tunnuseks, kuna tegu on dimensionaalses tekstimudelis afektiivsuse dimensiooniga ja seega afektiivsus ja formaalsus ei oleks iseseisvad dimensioonid. Propositsioonide arv mittespontaansuse tunnusena seevastu väljendab otseselt informatsioonitihedust, mis on omakorda iseseisev dimensioon. (Vaik jt 2020)

Tunnuste alamhulga valimisel otsustasin, et ei käsitle tunnuseid, mis vajavad süntaksianalüsaatorit, kuna süntaksipõhiste tunnuste realiseerimine ja analüüsimine ei mahtunud ajaliselt bakalaureusetöö raamidesse. Toorlaene ei vaata, kuna keeletuvastajad ei toimi väga edukalt üksikute sõnade keele tuvastamisel.

Ei vaata tunnusena ka sõnavara, kuna peamine allikas, kust saaks stiili alusel võtta sõnavara, on Eesti Keele Instituudi ühendsõnastik (Ühendsõnastik 2021). Ühendsõnastikus on stiili märgendid sõnadele märgendatud ka siis, kui ainult üks sõna tähendus on selle stiiliga, näiteks saavad kõnekeele märgendi ka sõnad *aamen* ja *kapsas*. Loendi käsitsi puhastamine on väga ajamahukas töö, mis oleks ületanud bakalaureusetöö mahu. Ei valinud tunnuste valimisse ka kirjavahemärkide puudumist ja sõnade kordumist, kuna nende realiseerimisel tekkis probleeme.

Tunnuste valimisse, mida kasutan dimensiooni hindamisel, valisin lõpuks 14 tunnust. Esiteks tunnused, mida vaatan ainult morfoloogilist analüüsi kasutades: sõnavara mitmekesisus vaadates lemmasid, nimisõnafraaside pikkus ja arv, lemmade keskmine pikkus, isikulised asesõnad ja verbid, impersonaal, *nud*-partitsiip, kaudne kõneviis. Teiseks leksikonil ehk loenditel põhinevad tunnused: emotikonid, tajuverbide osaarv. Kolmandaks vaatan reguraalavaldiste abil puuduva tühikuga kirjavahemärke ning sõnasiseseid korduseid. Neljandaks vaatan kirjavigadega sõnu ning sõnu, millel on kas puuduv suur algustäht või läbivalt suured tähed.

3. Meetod

3.1. Programmeerimiskeskkond

Töö käigus loodud automaatse hindamise süsteem on kättesaadav leheküljel https://github.com/kgailit/Gailit-2021. Leheküljel on kirjas, kuidas süsteemi kasutada ja link kasutatud alamkorpusele (Ühendkoprus 2019).

Programmi koostasin kasutades programmeerimiskeelt Python. Programmi kirjutasin keskkonnas Jupyter Notebook, kuid koostasin Notebookidest lõpliku kasutatava Pythoni skripti. Kasutasin ka Pythoni teeki EstNLTK, et tekste morfoloogiliselt analüüsida.

3.2. Andmestik

Skooride hindamise katsetamiseks kasutan eesti keele ühendkorpuse 2019. aasta veebitekstide alamkorpust. (Ühendkorpus 2019) Kasutasin eeltöötlemata varianti, mis on ainult lausestatud, et hinnatavad tekstid oleksid vormilt võimalikult sarnased veebis leiduvate tekstidega.

Pakitud korpuse pakkisin lahti eraldi tekstideks, millest koostada alamhulk. Alamhulga suuruseks otsustasin võtta 100 000 faili, kuna see on piisavalt suur hulk, et saada statistilist informatsiooni, kuid mida saaks ka käsitsi uurida. Valisin tekstide alamhulga suvaliselt, kuid jälgisin, et võimalikult palju allikaid oleks esindatud. Tekste alamhulka valides valisin esmalt suvalise allika ja seejärel allikast suvalise teksti. Tekste valides jälgisin, et nende pikkus oleks vähemalt 500 tähemärki, kaasa arvatud tühikud ja kirjavahemärgid. Tekstid, mis olid pikemad kui 200 lauset, lõikasin lühemaks, jättes alamhulga tekstiks ainult esimesed 200 lauset.

3.3. Hindamine

Dimensiooni skoori hindamiseks lõin süsteemi, mis võtab sisse teksti ja tagastab formaalsuse ja spontaansuse dimensioonide skoorid. Skoorid arvutatakse keeleliste ja tekstiliste tunnuste alusel. Esimene samm on tunnuste tuvastamine. Selleks kasutan EstNLTK (EstNLTK) keeletöötlusteeki, et tekst esmalt sõnestada ja seejärel teha morfoloogiline analüüs ehk morfanalüüs. Luuakse kaks EstNLTK Text-tüüpi objekti, kus ühes teostatakse morfoloogiline analüüs koos tundmatute sõnade analüüsi oletamisega ja teine ilma oletamiseta. Oletamise puhul saab iga sõna endale morfanalüüsi, oletamiseta morfanalüüsil on väljundis tundmatu sõna analüüsiga sõnad, mille algvormi ja grammatilisi kategooriad ei teata. Tundmatu sõna analüüsiga sõnad viitavad, et tegu võib olla sõnaga, mis sisaldab dimensiooni skoori määramiseks olulisi tunnuseid, näiteks kirjavigu või üldkeeles vähe levinud ja seega dimensioonile omaseid sõnu. Teine samm on tunnuste eraldamine sisendtekstist, mida kirjeldan iga tunnuse jaoks eraldi allpool. Kolmanda sammuna teisendan tunnused skooriks ning neljas samm on skoori normaliseerimine.

3.4. Tunnuste tuvastamise meetodid

3.4.1. Emotikonid

Emotikonid on süsteemis esimene tunnus, mida vaatlen. Seda teen juba töövoo esimese sammu juures, kus loon EstNLTK Text-objekte. Otsin algtekstist üles emotikonid ja emojid, teen nendest loendid ja eemaldan leitud emotikonid tekstist. Eemaldamine on vajalik, et sõnestamisel kirjavahemärkide tekstist lahku löömisel ei tekiks ühetähemärgilisi sõnesid, mis vähendaksid

keskmist sõnapikkust ja tekitaksid teisigi probleeme. Emotikonideta tekst võetakse EstNLTK Text-objektide sisendtekstiks.

Emotikone ja emojisid otsin kolmes jaos. Esimene jagu on regulaaravaldiste abil kahe kooloni vahele märgitud emotikonide otsimine. Vaatan, et tegu ei oleks valepositiivse tulemusega ehk tulemusega, mis näiliselt vastab nõutele, kuid tegelikult pole tegu emotikoniga vaid on mingil muul põhjusel samamoodi kahe kooloni vaheline tähekombinatsioon, näiteks lingisised *:http:* ja *:https:* märgenditega. Teiseks eraldan mitmemärgilised emotikonid, milles on vähemalt üks täht, mitte ainult kirjavahemärgid. Teen seda eraldi, et arvestada sõnadega kokku kleepunud kirjavahemärkidega, sest muidu kaoksid osadel sõnadel algused või lõpud ära. Kolmandaks ja viimaseks eraldanud kõik ülejäänud emotikonid, nii mitmetähemärgilised kui ka Unicode'i standardis ühemärgilised emojid.

Emotikonide tunnuseks on tekstist leitud emotikonide arv.

3.4.2. Sõnavara mitmekesisus

Sõnavara mitmekesisuse arvutan teksti lemmasid vaadates. Kõikide unikaalsete lemmade arvu jagan kõikide lemmade arvuga ja tunnuseks on saadud arv.

3.4.3. Nimisõnade ja nende laiendite arv

Nimisõnade ja nende laiendite arvu vaatan sõnaliikide abil, otsides nimisõnu, omadussõnu, pärisnimesid, kaassõnu ja arvsõnu. Võrdlen nende arvu kõigi sõnade arvuga ning arvutan nende protsendi.

3.4.4. Lemmade keskmine pikkus

Lemmade keskmist pikkust vaatan kasutades EstNLTK teeki sõnaliigi oletamisega. Liitsõnu vaatlen kasutades EstNLTK liitsõnaanalüsaatorit, jaotades liitsõna osalemmadeks ning käsitledes neid edasi kui lihtsõnade lemmasid. Tunnuseks arvutan keskmise lemma pikkuse liites kõikide leitud lemmade pikkused kokku ja jagades selle lemmade arvuga.

3.4.5. Tajuverbide osaarv

Tajuverbide osaarvu leian jagades tekstis leiduvate tajuverbide arvu kõigi tekstis leiduvate verbide arvuga. Tajuverbide loend on võetud eesti WordNetist võttes hüperonüümiks *tajuma.v.04* ehk verbi *tajuma* neljanda tähenduse WordNetis (WordNet). Selle sõna kõikide hüponüümide loendist olen eemaldanud ebasobivad verbid, nagu mitmest sõnast koosnevad verbid ning mitmetähenduslikud verbid, mille levinuim tähendus ei ole tajumisega seotud.

3.4.6. Lühikesed morfoloogilisel analüüsil tundmatuks jäänud sõnad

Lühikesi morfoloogilisel analüüsil tundmatuks jäänud sõnu vaatan kasutades EstNLTK sõnaliikide oletamiseta Text-objektide analüüse. Oletamiseta sõnade nimekirjast vaatan kõiki sõnu ning arvestan, et kui tekstis algab sõna suure algustähega, on tegu pärisnimega.

Lisaks vaatan, et sõna oleks lühike. Lühikeseks pean sõnu, mille pikkus tekstis on maksimaalselt 10 tähemärki. Selle piirarvu arvutan võttes aluseks keskmise lemma pikkuse korpuses, milleks on 4,83. Kuna morfoloogilisel analüüsil tundmatuks jäänud sõnad ei ole lemmad, liidan käände- ja pöördelõppude jaoks viis tähemärki. Ümardan piirarvu üles, saades võrreldavaks arvuks 10. Lühikeste tundmatu sõnaliigi analüüsiga sõnade protsendi arvutan jagades nende arvu kõikide sõnade arvuga.

3.4.7. Puuduva tühikuga kirjavahemärgid

Puuduva tühikuga kirjavahemärke otsin kasutades regulaaravaldist. Puuduva tühikuga kirjavahemärkidena käsitlen vaid juhtumeid, kus kirjavahemärk, nagu koma või koolon, on tühikuta kahe sõna vahel. Ma ei käsitle sidekriipsu kui kirjavahemärki, sest seda leidub ka õigesti kirjutatud sõnades. Tunnus on arv kohtadest, kus kirjavahemärk on jäänud tühikuta.

3.4.8. Isikulised asesõnad ja verbid

Isikulisi asesõnu vaatan kasutades EstNLTK morfanalüsaatorit. Kui sõna saab asesõna märgendi, vaatan sõna lemmat. Kui see on üks kolmest isikulisest asesõnast (*mina*, *sina*, *tema*), suurendan selle esinemiste arvu. Mitmuslikke asesõnu (*meie*, *teie*, *nemad*) käsitleb EstNLTK kui ainsusliku vormi mitmust, ehk asesõna lemma on alati ainsuslik asesõna. Teise isiku formaalne ainsuslik

asesõna *Teie* saab seega samuti lemmaks *sina*, seetõttu pean teietamist ainult teise isiku kasutamiseks. Tunnuseks on iga isikulise asesõna protsent kõikidest asesõnadest, mitte ainult isikulistest.

Verbide isikuid vaatan kasutades EstNLTK morfanalüsaatorit. Kui sõna saab tegusõna märgendi, vaatan sõna vormi analüüsi. EstNLTK tagastab vormi analüüsi lõpuformatiivina, mille alusel teen isikuliste verbivormide loendid isikute kaupa. Kui lõpuformatiiv võib tähistada mitut isikut või kui verbi analüüs jääb üleüldiselt mitmeseks, ei käsitle ma seda kui kindla isiku tunnust. Tunnuseks on iga verbi isiku (esimene, teine ja kolmas isik) kohta protsent, kui palju on selles isikus verbe võrreldes kõikide verbidega tekstis.

3.4.9. Impersonaal

Impersonaali vaatan kasutades EstNLTK morfanalüsaatorit. Kui sõna saab tegusõna märgendi, vaatan morfanalüüsist verbivormi järgi, kas sõna on impersonaalne. Kui verbivorm väljendab impersonaali, suurendan impersonaalsete verbide arvu ühe võrra. Kui EstNLTK ei suuda morfoloogilist analüüsi ühestada ja analüüs jääb mitmeseks, vaatan kõiki analüüse ning kui vähemalt üks analüüs on impersonaalne, pean verbi impersonaalseks. Tunnuseks on impersonaalsete verbide protsent kõikidest verbidest.

3.4.10. nud-partitsiip

nud-partitsiipi vaatan kasutades EstNLTK morfanalüsaatorit. Kui sõna saab tegusõna märgendi, vaatan, et see oleks saanud *nud*-partitsiibi analüüsi ja suurendan tekstis esinemise arvu ühe võrra. Kui EstNLTK ei suuda morfoloogilise analüüsi väljundit ühestada ja analüüs jääb mitmeseks, vaatan kõiki analüüse ning kui vähemalt üks analüüs on *nud*-partitsiibi märgendiga, pean verbi *nud*-partitsiibiks. Tunnuseks on *nud*-partitsiibi protsent kõikidest verbidest.

3.4.11. Kaudne kõneviis

Kaudset kõneviisi käsitlen ainult *vat*-tunnuselise kaudse kõneviisina. Kasutades EstNLTK morfanalüsaatorit vaatan, et kui sõna saab tegusõna sõnaliigi ja kaudse kõneviisi märgendi, siis suurendan kaudse kõneviisi tekstis esinemiste arvu ühe võrra. Kui EstNLTK ei suuda morfoloogilise analüüsi väljundit ühestada ja analüüs jääb mitmeseks, vaatan kõiki analüüse ning

kui vähemalt üks analüüs on kaudse kõneviisi märgendiga, pean verbi kaudseks. Tunnuseks on kaudse kõneviisi protsent kõikidest verbidest.

3.4.12. Puuduv suur algustäht

Puuduvaid suuri algustähti otsin ainult lausete ja pärisnimede algustest. Teksti jagan lauseteks kasutades EstNLTK lausestajat. Esmalt vaatan iga lause esimest tähte: kui see väike, suurendan puuduvate suurte algustähtede arvu ühe võrra. Tunnuseks on protsent, kui paljud sõnad on tekstis puuduvate suurte algustähtedega.

3.4.13. Läbivalt suurte tähtedega sõna

Sõnu, mis esinevad tekstis läbivalt suurte tähtedega, otsin tekstist vaadates iga sõna puhul, et sõna oleks vähemalt kaks tähemärki pikk ning et kõik selle tähed oleksid suured. Kontrollin, et EstNLTK ei ole andnud sõnale lühendi analüüsi, sest siis ei oleks tegu tavalise sõnaga, mis on kirjutatud läbivate suurtähtedega. Kui sõna on läbivalt suurtähtedega ja ei ole lühendi analüüsiga, suurendan läbivalt suurte tähtedega sõnade arvu ühe võrra. Tunnuseks on protsent, kui paljud sõnad on tekstis läbivalt suurte tähtedega.

3.4.14. Sõnasisesed kordused

Sõnasiseseid korduseid otsin regulaaravaldise abil. Vaatan eraldi sõnade sees leiduvaid ühe tähe kordusi ja mitme tähe pikkuseid korduseid. Ühe tähe korduste puhul vaatan, et täht korduks sõnas vähemalt neli korda. Vähemalt kahe tähe pikkuste korduvate tähekombinatsioonide puhul vaatan, et need korduksid kolm või enam korda. Ei vaata väiksemaid kordusi, kuna need võivad esineda grammatiliselt korrektses eesti keeles. Loetlen sõnasisesed kordused kokku ja nende arv on tunnuseks.

3.4.15. Kirjavead

Kirjavigade otsimiseks kasutan EstNLTK teeki sisse ehitatud kirjavigade analüüsi. Vaatan iga tekstis esinevat sõna järjest. Kui sõna on vigaseks märgitud, vaatan, et tegu ei oleks pärisnime analüüsi saanud sõnaga, sest need on tihti valesti kirjavigasteks märgendatud. Kui tegu ei ole

pärisnimega, suurendan kirjavigaste sõnade arvu ühe võrra. Tunnuseks on kirjavigadega sõnade protsent kõikidest sõnadest tekstis.

3.5. Tunnuste punktideks teisendamine

3.5.1. Kirjeldus

Tekstist eraldatud tunnused teisendan punktideks kasutades korpuse põhjal arvutatud võrdlusi. Olenevalt tunnusest, mõjutab mõni tunnus dimensiooni skoori positiivselt ja mõni tunnus negatiivselt. Tabelites olen selle välja toonud eraldi tulbana. Teisendused jagunevad kolmeks teisendustüübiks: binaarne, ühesuunaline ja kahesuunaline.

Binaarseteks tunnusteks pean tunnuseid, mida esines alla poolte korpuses olevatest tekstidest. Binaarsete teisenduste puhul vaatan, kas tunnus tekstis leidub, ning kui leidub, annan tunnuse maksimaalse punktide arvu, vastasel juhul annan null punkti.

Ühesuunalisi ja kahesuunalisi teisendusi vaatan skaalal, mille olen arvutanud korpuse põhjal. Kui tekstis esineb tunnust samal määral, nagu seda esineb korpuses kõige levinumalt, saab tekst tunnuse eest null punkti. Mida rohkem erineb tunnuse esinemine korpuse keskmisest, seda rohkem punkte antakse või võetakse.

Ühesuunaliste ja kahesuunaliste teisenduste erinevus on, et ühesuunalised teisendused saavad punkte, kas ainult juurde anda või ära võtta, kuid kahesuunalised võivad teha mõlemat, olenevalt, kuidas tunnus tekstis esineb. Tabelites olen selguse huvides jaganud kahesuunalised tunnused pooleks, ehk näitan eraldi tunnuse positiivset ja negatiivset mõju punktidele.

Ühesuunaliste tunnuste puhul olen jaganud kogu korpuse tunnused väärtuste alusel kuueks võrdseks grupiks. Esimene kuuendik on kõige väiksemad väärtused, viimane kuuendik on kõige suuremad väärtused. Mida väiksem väärtus, seda vähem tunnus tekstis esineb. Seega, kui tekst kuulub tunnuse poolest esimesse kuuendikku, saab tekst tunnuse eest null punkti, teises kuuendikus ühe punkti, kolmandas kaks punkti, neljandas kolm punkti, viiendas neli punkti ning kuuendas ehk viimases kuuendikus viis punkti.

Kahesuunaliste tunnuste puhul olen jaganud korpuse väärtused üheteistkümneks võrdseks grupiks. Sarnaselt ühesuunaliste tunnustega, on esimene üheteistkümnendik kõige väiksemad tunnused ja viimane üheteistkümnendik kõige suuremad tunnused.

Skooride jaoks punktide andmisel vaatan, et ühe tunnuse eest ei antaks mitu korda punkte. Seda pean jälgima kahes kohas: isikuliste verbide ja asesõnade puhul ning kirjavigadega sõnade vaatamisel. Isikulisi verbe ja asesõnu vaadates on formaalsuse tunnuseks kolmanda isiku kasutamine ning mitteformaalsuse tunnuseks esimese ja teise isiku kasutamine. Kirjavigadega sõnu vaatan kahes kohas, nii EstNLTK kirjavigade tuvastamist kasutades kui ka morfanalüüsil tundmatu sõna sõnaliigiga sõnu vaadates. Kõigil kolmel juhul liidan seotud tunnuste eest antud punktid kokku ja jagan saadud punktisumma seotud tunnuste arvuga, enne dimensiooni punktisummaga liitmist.

3.5.2. Teisendustabelid

Tunnuste skoorideks teisendamisel vaatasin korpuses tunnuste esinemisi. Võrdlen vaadeldava teksti tunnust korpuse tunnustega ning annan vastavalt punkte. Punktide andmise aluseks olevad korpuse tunnuste väärtused olen pannud tabelitesse, kus on näha, millised tunnuste väärtuste vahemikud annavad millise arvu punkte. Tabelites olen loetavuse jaoks ümardanud tunnuste väärtused kolme komakohani. > tähistab, et tunnusel seda punktivahemikku ei vaadata ning antakse parempoolse tulba punktid.

Formaalsuse tabel

Tunnus	Mõju	0 punkti	1 punkt	2 punkti	3 punkti	4 punkti	5 punkti
Sõnavara mitmekesisus	+	0.647 – 0.680	0.680 – 0.710	0.710 – 0.740	0.740 – 0.773	0.773 – 0.813	0.813 – 1
Sõnavara mitmekesisus	-	0.680 – 0.647	0.647 – 0.612	0.612 – 0.571	0.571– 0.520	0.520 – 0.450	0.450 – 0
Keskmine lemmapikkus	+	4.950 – 4.855	4.855 – 4.761	4.761 – 4.658	4.658 – 4.544	4.544 – 4.402	4.402 – 0
Keskmine lemmapikkus	-	4.855 – 4.950	4.950 – 5.046	5.046 – 5.154	5.154 – 5.282	5.282 – 5.469	5.469 – 1

Esimese isiku asesõnad[1]	-	0	0 – 0.242	0.242 – 0.500	0.500 – 0.667	0.667 – 0.895	0.895 – 1
Teise isiku asesõnad[1]	-	0	>	>	0 – 0.200	0.200 – 0.500	0.500 – 1
Esimese isiku verbid[1]	-	0	>	0 – 0.105	0.105 – 0.222	0.222 – 0.375	0.375 – 1
Teise isiku verbid[1]	-	0	0 – 0.059	0.059 – 0.125	0.125 – 0.200	0.200 – 0.333	0.333 – 1
Kolmanda isiku asesõnad[2]	+	0	>	0 – 0.200	0.200 – 0.417	0.417 – 0.800	0.800 – 1
Kolmanda isiku verbid[2]	+	0 – 0.400	0.400 – 0.537	0.537 – 0.667	0.667 – 0.800	0.800 – 0.946	0.946 – 1
Emotikonid	-	0	>	>	>	>	0 < ...
Kaudne kõneviis	+	0	>	>	>	>	0 – 1
nud-partitsiip	-	0	0 – 0.011	0.011 – 0.031	0.031 – 0.049	0.049 – 0.074	0.074 – 1
Impersonaal	+	0 – 0.020	0.020 – 0.043	0.043 – 0.065	0.065 – 0.096	0.096 – 0.148	0.148 – 1
Nimisõnafraasid	-	0.503 – 0.476	0.476 – 0.447	0.447 – 0.415	0.415 – 0.379	0.379 – 0.333	0.333 – 0
Nimisõnafraasid	+	0.476 – 0.503	0.503 – 0.528	0.528 – 0.553	0.553 – 0.581	0.581 – 0.620	0.620 – 1

Tabel 1: Formaalsuse tunnuste teisendamine punktideks.

[1] tunnuseks on esimese ja teise isiku kasutamine, seega et tunnus ei mõjutaks formaalsuse hinnangut teistest tunnustest rohkem, liidan esimese ja teise isiku verbide ja aseõnade eest antud punktid kokku ja jagan neljaga, et see annaks maksimaalselt sama palju punkte, kui teised tunnused.

[2] tunnuseks on kolmanda isiku kasutamine, seega et tunnus ei mõjutaks formaalsuse hinnangut teistest tunnustest rohkem, liidan kolmanda isiku verbide ja aseõnade eest antud punktid kokku ja jagan kahega, et see annaks maksimaalselt sama palju punkte, kui teised tunnused.

Spontaansuse tabel

Tunnus	Mõju	0 punkti	1 punkt	2 punkti	3 punkti	4 punkti	5 punkti
Sõnavara mitmekesisus	+	0.680 – 0.647	0.647 – 0.612	0.612 – 0.571	0.571– 0.520	0.520 – 0.450	0.450 – 0
Sõnavara mitmekesisus	-	0.647 – 0.680	0.680 – 0.710	0.710 – 0.740	0.740 – 0.773	0.773 – 0.813	0.813 – 1
Emotikonid	+	0	>	>	>	>	0 < ...
Kirjavigadega sõnad[3]	+	0	0 – 0.008	0.008 – 0.013	0.013 – 0.021	0.021 – 0.035	0.035 – 1
Lühikesed tundmatu sõna analüüsiga sõnad[3]	+	0	0 – 0.002	0.002 – 0.007	0.007 – 0.012	0.012 – 0.022	0.022 – 1
Keskmine lemmapikkus	-	4.855 – 4.950	4.950 – 5.046	5.046 – 5.154	5.154 – 5.282	5.282 – 5.469	5.469 – 1
Keskmine lemmapikkus	+	4.950 – 4.855	4.855 – 4.761	4.761 – 4.658	4.658 – 4.544	4.544 – 4.402	4.402 – 0
Tajuverbid	+	0 – 0.044	0.044 – 0.086	0.086 – 0.118	0.118 – 0.150	0.150 – 0.193	0.193 – 1
Tühikuta kirjavahemärgid	+	0	>	>	>	>	0 < ...
Läbinisti suured sõnad	+	0	>	>	0 – 0.002	0.002 – 0.009	0.009 – 1
Puuduva suure algustähega	+	0	>	>	>	>	0 < ...
Esimese isiku asesõnad	+	0	0 – 0.242	0.242 – 0.500	0.500 – 0.667	0.667 – 0.895	0.895 – 1

Tabel 2: Spontaansuse tunnuste teisendamine punktideks.

[3] mõlemad, kirjavigade sõnade protsent ja lühikeste morfanalüüsil tundmatu sõna sõnaliigi saanud sõnade protsent, väljendavad tunnusena kirjavigaste sõnade esinemist, seega, et tunnus ei mõjutaks spontaansuse hinnangut teistest tunnustest rohkem, liidan nende eest antud punktid kokku ja jagan kahega, et see annaks maksimaalselt sama palju punkte, kui teised tunnused.

4. Evalveerimine

4.1. Inimhinnanguga võrdlemine

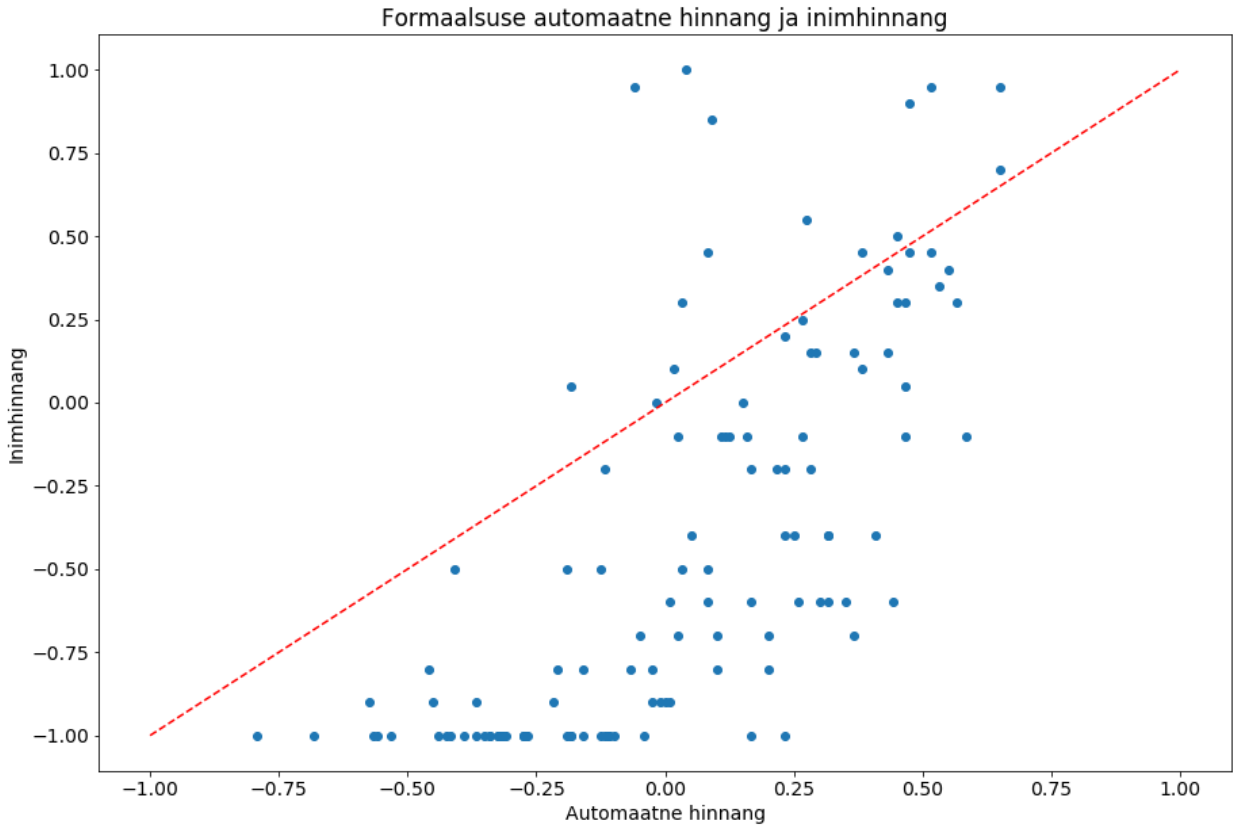
Dimensioonide automaatse hinnangu võrdlemiseks inimhinnangutega kasutan oma juhendajalt saadud spontaansuse ja formaalsuse suhtes käsitsi märgendatud testkorpust. Testkorpus koosneb 120 tekstist, mis pärinevad etTenTen13 korpusest (etTenTen). Annoteerijatel tuli igas hindamissessioonis hinnata tekste dimensioonide kolmikute kaupa. Annoteerijatele esitati korraga kaks teksti (tekst A ja tekst B) ning ülesanne oli valida vaatluse all olevale dimensioonile kõige iseloomulikum tekst (kas A või B). Seda iseloomulikkust hinnati 4-pallilisel järjestusskaalal: 1 (dimensiooni esineb nõrgalt), 2 (dimensiooni esineb mõõdukalt), 3 (dimensiooni esineb tugevalt) ja 0 (dimensioon puudub). Kui annoteerija valis teksti A, siis tekst B sai automaatselt vaatluse all olevas dimensioonis hinnanguks 0, ja vastupidi. Juhul, kui annoteerija ei suutnud kahe teksti vahel valida, said mõlemad tekstid automaatselt vaatluse all olevas dimensioonis hinnanguks 0.

Selle annoteerimiskatse tulemusena tekkis korpus, kus on esindatud tekstid, mis said kõigilt annoteerijatelt hinnanguks > 0, kuni sellisteni, mida ükski annoteerija ei valinud kordki ning mille keskmine hinnang on null. Märkusena peab mainima, et kui tekst saab omale keskmiseks hinnanguks 0, siis see ei pruugi automaatselt tähendada, et dimensiooni tekstis üldse ei esineks, vaid põhjuseid võib olla mitmeid, näiteks jäi ülesandepüstitus arusaamatuks, liiga vähe konteksti (tekstikatked olid liiga lühikesed), kahe võrdse teksti puhul ei suudetud otsustada (kuigi mõnes teises kontekstis võib üks tekst teatud dimensioonis valituks osutuda).

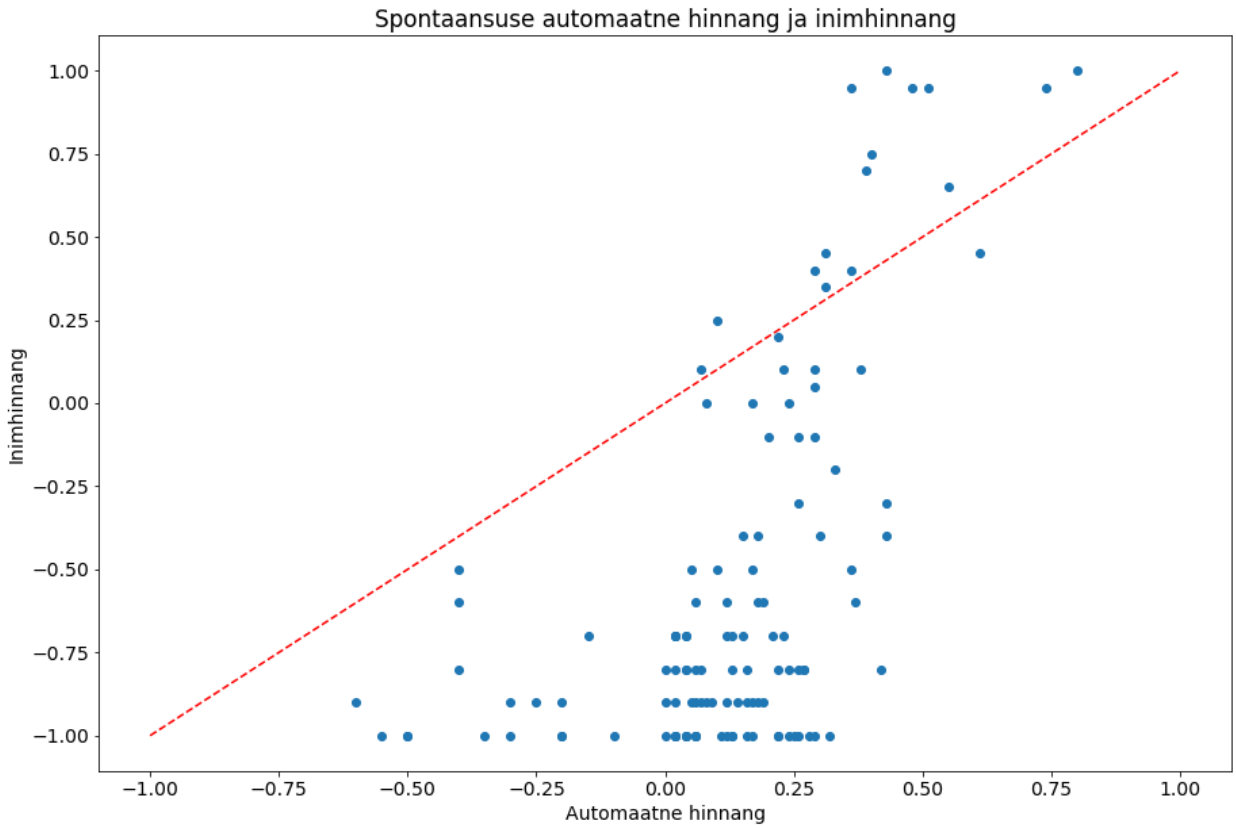
Testkorpuse võrdlemiseks automaatselt hinnatud korpusega kasutan testkorpuses olevaid eelnevalt arvutatud hinnangute keskmisi ning teisendan need vahemikku -1–1, mida on kasutatud automaatsel hindamisel. Et saaksin testkorpuse hinnanguid võrrelda automaatse hindamisega, eeldan, et kui tekstile on antud hinnanguks 0, siis on tegu negatiivse hinnanguga, nagu automaatse hindamise skoor -1. Teisendamisel jälgin, et positiivsed hinnangud jääksid positiivseteks ja negatiivsed negatiivseteks, seega ei saa lihtsalt teisendada vahemikku 0–3 vahemikuks -1–1. Hinnangud vahemikus 0–1 on teisendatud vahemikku -1–0, lahutades inimeste antud keskmisest hinnangust ühe. Hinnangud vahemikus 1–3 on teisendatud vahemikku 0–1, lahutades esmalt inimeste antud keskmisest hinnangust ühe ja seejärel jagades saadud vahe kahega. Kui inimesed

on andnud tekstile keskmiseks hinnanguks 1, on selle teisendatud skoor 0, seega 1 on neutraalne hinnang.

Automaatse hindamise võrdlemiseks inimhinnanguga tegin esmalt joonised, et näha üldiselt, kuidas hinnangud erinevad. Joonis 1 võrdleb formaalsuse automaatset hinnangut inimhinnangutega, joonis 2 võrdleb spontaansuse hindamisi. Mõlemal joonisel märgib punane punktiirjoon ühisosa, ehk punasel joonel on automaatne hinnang ja inimhinnang sama. Mõlemal joonisel on näha, et automaatne hinnang ei lähe nii palju ekstreemumitesse, kui inimhinnang, kuid see tuleb palju paremini esile spontaansuse puhul. Lisaks on inimhinnang palju madalama keskmisega, kui automaatne hinnang.



Joonis 1: Formaalsuse automaatse hinnangu ja inimhinnangu võrdlus.



Joonis 2: Spontaansuse automaatse hinnangu ja inimhinnangu võrdlus.

4.1.1. Formaalsuse hindamine

Mitteformaalne (inimeste hinnang) vs formaalne (automaatne hindamine)

Tekstides, mida inimesed on hinnanud mitteformaalsetena ja mida on automaatselt formaalseteks märgitud, on näha, et leidub mitmeid tunnuseid, mis põhjustavad formaalsuse erinemist inim- ja automaatse hinnangu vahel. On näha, kuidas sõnavara ja lausestruktuur mõjutavad inimeste formaalsuse hindamist. Näide 1 on pärit Pealinna ajalehe veebileheküljelt. Selle formaalsust on inimesed hinnanud skooriga -0.6 ja automaatselt on seda hinnatult skooriga 0.442:

> (1) *EE juhatuse nõudmisel on miljarditesse kroonidesse ulatuvate kulutuste tegemiseks raha vaja nüüd ja kohe. Muidu jäävat, tulenevalt vanade energiaplokkide sulgemise vajadusest, Eesti aastast 2016 elektrita. Samas aga kulutatakse just praegu umbes 1,5 miljardit krooni olemasolevatele vanadele energiaplokkidele väävlipuhastuse seadmete soetamiseks eesmärgiga tagada nende plokkide energiavarustuskindlus võimsusega 700*

MW aastani 2025. Ehk siis praegusega võrreldav aastane tootmismaht 8700 GWh. Lisaks on kasutuses uued CFB energiaplokid võimsusega 430MW. 2013. aastal valmib ka uus 650 MW Estlink 2 kaabel Põhjamaade energiasüsteemiga täiendava ülekandevõimsuse tagamiseks. Seega, kas Eestit ikka ähvardab elektripuudus, kui EE kohe miljardeid kroone ei saa? Eesti tarbijat elektripuudus ei ohusta Eesti tarbijate aasta keskmine tarbimiskoormus on enam kui tagatud. Lühiajalise tippkoormuse katmise tagab energiaühendus Põhjamaadega, Venemaaga, Läti-Leeduga. Nii toimib elektriühisturg kõigis Nord Pool liikmesriikides. Mitte ükski neist riikidest ei taga aasta ringi 100% enda tarbimisvajadust vaid enda jaamade toodetud elektriga. Sest majanduslikult pole niiviisi lihtsalt ratsionaalne. See ongi Nord Pool süsteemi mõte. Arvestame ka seda, et juba täna lahkuvad paljud suurtarbijatest kliendid Eesti Energia tegevusest tulenevalt hoopis konkurentide juurde. Sealhulgas Eesti Energia nõukogu liikmed!!! Eelnevast lähtuvalt on selgusetu, miks on EE-l kiiresti ja kohe vaja saada raha kahe uue põlevkiviploki hankeks. (etTenTen www.pealinn.ee, doc id = 132426)

Tekstis on mitmeid mitteformaalseid sõnu, fraase ja muid tähiseid, nagu *100%* ja *!!!*, mis teevad teksti lugejate jaoks vähem formaalseks ja rohkem argiseks. Seega kuigi tekstil on formaalsuse tunnuseid (sõnavara mitmekesisus 0.73, kaudse kõneviisi kasutamine), võib sõnavara ja lauseehitus muuta selle inimeste jaoks mitteformaalseks.

Sõnavara ja lausestruktuuri lisamine tunnusteks suurendaks ka mitmete praeguse hindamise järgi vähesel määral mitteformaalsete tekstide mitteformaalsust. Näide 2 pärineb ajalehe Eesti Ekspress veebilehe kommentaaridest. Inimesed on selle formaalsust hinnanud skooriga -1 ja automaatselt on seda hinnatud -0.117:

(2) *? 22.04.2010 09:54*

- Priit Hõbemägi: Pevkuril digiretsepti ei ole. Sinul on. Talle see meeldib. Vasta

ahah 22.04.2010 12:50

- Tavaline dumbjuuserite käitumine: kui leht kohe kohale ei tule, siis vajuta veelkord refreshi- äkki tuleb? Kui sel korral kah ei tulnud, siis mida teeb dumbjuuser? MUIDUGI! ELEMENTAARNE- TA VAJUTAB VEEL PAAR KORDA, IGAKS JUHUKS!!!111111ELEVEn idioodid Vasta

tohoh 22.04.2010 13:20

- ahah juuser või dumbjuuser, selline asi peab olema igale normaalse intelligentsiga kodanikule lihtsalt kasutatav, seda, et IT-mehed oskavad kasutada, pole keegi vaidlustanud apteegid on juba kümmekond aastat kasutanud müügi-laoarvestus-retseptiprogramme ja digiretsepti oma ei tohi olla neid keerulisem, vastupidi, peab olema lollikindel ja ega ta polegi eriliselt raske, selle probleemid on retseptikeskuses mitte refreshivõimelistes apteekides retseptikeskus ei asu apteegis, et sinusugune juuser ikka aru saaks

Riho Kurg 22.04.2010 13:48

- Eestis on olemas inimesed ja oskused. Tehakse süsteeme, mis vastavad 200000 päringut (Swedbank) ja mis teenindavad 200 miljonit kasutajat (Skype). Kogu hiljuti toimunud Garage48 üritus oleks olnud suhteliselt mõttetu, kui teenusepakkuja HAVirtual-i võimekus oleks Microlinki klassis. Kas antud juhul oli tegu veaga analüüsifaasis, arendajate/haldajate ebakompetentsusega või lihtsa vargusega, ei oska ilmselt keegi vastata. Tegelikult ei tahaks uskuda, et sotsmin selle jama hostimise eest 700KEEK kuus maksab, loodetavasti on tegu ajakirjandusliku liialdusega. Vasta (etTenTen www.ekspress.ee, doc id = 495829)

Peab täheldama, et näites 2 on ka mitmeid spontaansuse tunnuseid, mida ma ei loe mitteformaalsuse tunnusteks. Näiteks toodud tekst ei saanud automaatsel hindamisel madalamat skoori, kuna sellel on mitmeid tunnuseid, mis on formaalsusele omased, nagu rohke kolmanda isiku ning *nud*-partitsiibi kasutamine, seega võib eeldada, et need kaks tunnust võivad mõjutada dimensiooni skoori liiga palju.

Formaalse (inimeste hinnang) vs mitteformaalne (automaatne hindamine)

Leidub ka tekste, mida inimesed on hinnanud formaalseteks ja mida on automaatselt hinnatud vähe- või mitteformaalseteks. Näide 3 pärineb veebilehelt no.spam.ee ja selle inimhinnanguline formaalsus on 0.95 ja automaatne formaalsus -0.058:

> (3) *Andmekaitse Inspektsioonile laekus 15.07.2010 Margus Lehesaare kaebus Teie tegevuse peale seoses tema isikuandmete avalikustamisega võrgulehel http://no.spam.ee/?p=236. Margus Lehesaare väitel ei ole ta andnud Teile nõusolekut oma isikuandmete avaldamiseks eeltoodud võrgulehel. Lähtudes eeltoodust on Andmekaitse Inspektsioon alustanud isikuandmete kaitse seaduse § 32 lõike 1 ning haldusmenetluse seaduse § 35 alusel riiklikku järelevalvemenetlust. Isikuandmete kaitse seaduse § 1 sätestab, et isikuandmete kaitse seaduse eesmärk on kaitsta isikuandmete töötlemisel füüsilise isiku põhiõigusi ja -vabadusi, eelkõige õigust eraelu puutumatusele. Eraelu puutumatuse osana on igaühel põhiõigus isikuandmete kaitsele. Eraellu sekkub andmete kogumine, salvestamine, säilitamine, muutmine ja kasutamine, sh väljastamine kolmandatele isikutele ja avalikustamine ehk igasugune isikuandmete töötlemine. Isikuandmete kaitse seaduse § 10 lõike 1 kohaselt on isikuandmete töötlemine lubatud üksnes inimese nõusolekul, kui seadus ei sätesta teisiti. Kui andmed on saadud avalikest allikatest, ei tähenda see seda, et ilma andmesubjekti nõusolekuta võiks neid igal pool piiramatult töödelda. Nii on ka antud juhul Margus Lehesaarel õigus nõuda isikuandmete kaitse seaduse § 12 lõike 4 alusel oma isikuandmete avalikustamise lõpetamist võrgulehel http://no.spam.ee/?p=236.* (etTenTen no.spam.ee, doc id = 2227)

Kuigi näide 3 ise on formaalne, ei ole hindamiseks valitud tunnuste kaudu see välja tulnud. Ka selles näites on lausestruktuur ja sõnavara mõjutanud formaalsust, tehes seekord teksti formaalsemaks. Mitteformaalne automaatne hinnang tuleneb madalast teksti sõnavara varieerumisest, 0.567, ning asjaolust, et pooled teksti verbid ja asesõnad on teises isikus.

Mitteformaalse (inimeste hinnang) vs neutraalne (automaatne hindamine)

Tekstide hulgas, mida inimesed on hinnanud mitteformaalseks, kuid mida on automaatselt tunnuste alusel hinnatud neutraalselt, on näha, et inimesed on hinnanud mitteformaalseid ning neutraalsema formaalsusega tekste samamoodi. See tähendab, et tekstiliigid nagu intervjuud ja

retseptid saavad sarnasema formaalsuse hinnangu solvavate netikommentaaridega kui ajaleheartiklitega. Näide 4 on üks Nami-Nami retsept. Sellele on inimesed andnud hinnangu -0.9 ja automaatselt antud hinnang on -0.008.

> (4) *Lõika alõtša-ploomid pooleks ja eemalda kivid. Pane koos veega potti ja kuumuta keemiseni. Keeda tasasel tulel kaanega potis 15 minutit, kuni ploomid on pehmed. Tambi uhmris koriandriseemned, apteegitilliseemned, hakitud küüslauk, Cayenne ja sool ühtlaseks pastaks. Kui ploomid on pehmed, aja need läbi hakkmasina ja pane puhta poti sisse. Kuumuta keemiseni ning keeda mõõdukal tulel pidevalt segades umbes 3 minutit. Lisa vürtsipasta ja keeda veel umbes 5 minutit, kuni pasta pakseneb kergelt. Lisa hakitud münt ja koriander ja tõsta pott tulelt. Kalla tkemali kuumalt purkidesse. Jahuta toatemperatuurini ja siis säilita külmkapis. Kui soovid tkemalit kauem säilitada, siis kaaneta purgid õhukindlalt. Pildi tegi VIKE.* (etTenTen www.nami-nami.ee, doc id = 173054)

Kuigi retseptid ei ole formaalne tekstiliik, on tegu formaalsuse suhtes pigem ainult natukene, mitte väga, mitteformaalse liigiga, seda peamiselt lühiduse ja teise isiku ainsuse käskiva kõneviisi tõttu. Testkorpuses tähistab hindaja antud 0 aga, et tema arvates ei leidu dimensiooni tekstis. Seega testkorpuse hinnang -1 võib tähendada nii mitteformaalset kui ka neutraalse formaalsusega tekste, mitte ainult mitteformaalseid tekste, nagu seda tähendab automaatse hindamise -1. Leian, et kui varasem näide Eesti Ekspressi kommentaariumist, näide 2, on saanud inimestelt hinnangu -1, siis näide 4 ei ole piisavalt mitteformaalne, et saada hinnanguks -0.9.

4.1.2. Spontaansuse hindamine

Mittespontaanne (inimeste hinnang) vs neutraalne (automaatne hindamine)

Erinevalt formaalsusest, ei saa spontaansuse juures täheldada, et inimesed oleksid pidanud mittespontaanseteks tekstideks nii mittespontaanseid kui ka spontaansuse poolest neutraalseid tekste. Siiski, selliseid hindamisi leidus üksikuid. Näide 5 on võetud veebilehelt xn--eestimngula-q8a.ee. Selle inimhinnang on -1 ja automaatne hinnang 0.29:

(5) *Asume Tallinnas Haabersti linnaosas, Õismäe ja Mustamäe piiril, ärihoones aadressil Järveotsa tee 54. Paikneme Õismäe Vabaajakeskuse vahetus läheduses, Viimistluskeskuse majas, Melasi auto kohal teisel korrusel. Meie juures on parkimine tasuta ja kohti maja ees on ca 20-le autole. Lähiübruses veel paarikümnele autole. ÜHISTRANTSPORDIGA TULLES: Meie juurde saab Bussidega 16 ja 33, b ussipeatuse nimi : Autobussikoondis, tuleb tulla üle tee ja liikuda Õismäe suunas üle suure haljasala, Viimistluskeskuse maja jääb haljasalast vasakule. Eesti Mängulasse pääseb ka trollidega 6 ja 7, peatuse nimi: Nurmenuku, tuleb liikuda Rimist ja Nurmenuku turust paremalt mööda ja otse ees ongi Viimistluskeskuse maja.* (etTenTen www.xn--eestimngula-q8a.ee, doc id = 40079)

Näites on kasutatud läbivaid suurtähti (*ÜHISTRANTSPORDIGA TULLES*), kuid see on vähem emotsiooni või isikliku stiili väljendamiseks ja rohkem rõhutamiseks ning vormistatud veebilehel ei pruugi suurtähtede kasutus toimida seetõttu spontaanselt. Tekstis leidub mitmeid kirjavigu (*Lähiübruses*, *b ussipeatuse* ja *ÜHISTRANTSPORDIGA*), kuid neid pole märgitud kirjavigadeks. Morfanalüüsita jäänud lühikeste sõnade protsent on 0.88%. Lisaks on kõik kasutatud asesõnad esimese isiku asesõnad ning keskmine lemmapikkus on madal.

Mittespontaanne (inimeste hinnang) vs spontaanne (automaatne hindamine)

Tekstides, mida inimesed on hinnanud mittespontaansetena ja mida on automaatselt spontaanseteks märgitud, on tihti näha palju pärisnimesid. Võib eeldada, et neid ei ole EstNLTK alati pärisnimedeks märkinud ja seega käsitleb süsteem neid edasi kirjavigastena, kuid lühikeste morfanalüüsita sõnade vaatamisel ei vaadata pärisnimede tõttu suurtähega algavaid sõnu. Näide 6, mis pärineb leheküljelt bsd.ee, on üks selline tekst. Selle näite inimhinnang on -0.867 ja automaatne hinnang 0.27:

(6) *Esimene CDROM (ja üldine internetis levitatav) versioon oli FreeBSD 1.0, mis lasti välja 1993. aasta detsembris. See põhines 4.3BSD-Lite (``Net/2'') lindil U.C. Berkeleyst, koos mitmete muude komponentidega 386BSD ja Free Software Foundationi projektidest. Esimest versiooni saatis esimese väljaande kohta üsna hea edu ja tema järglaseks sai äärmiselt edukas FreeBSD 1.1 väljaanne 1994. aasta märtsis. Umbes sellel ajal moodustusid silmapiirile üsna ootamatud tormipilved, kui Novell ja U.C. Berkeley lahendasid oma pikaajalise kohtuprotsessi Berkeley Net/2 lindi seaduslikkuse üle. Selle*

protsessi üheks tingimuseks oli U.C. Berkely mööndus, et suur osa Net/2 koodist on ``piiratud'' kood ja kuulub Novellile, kes on selle omakorda omandanud AT&T käest. Vastutasuks andis Novell 4.4BSD-Lite'ile ``õnnistuse'' ja lubaduse, et kui 4.4BSD-Lite välja tuleb, saab ta olema ilma litsentsipiiranguteta ja kõigil Net/2 kasutajatel soovitatakse tungivalt sellele üle minna. (etTenTen www.bsd.ee, doc id = 638246)

Kuigi näide on mittespontaanne, ei ole see automaatselt hinnates välja tulnud. Näites 6 on nii lühikeste morfanalüüsil tundmatuks jäänud sõnade protsent 0,61% ning kirjavigadega sõnade protsent 4,27%, mis kinnitab eeldust, et pärisnimesid märgitakse kirjavigasteks. Tekstis leidub ka palju suurtähtlühendeid, mis ei ole endale saanud morfoloogilisel analüüsil lühendi märgendit: läbivalt suurte tähtedega sõnu on 5,97%. Lisaks teeb teksti spontaansemaks lühike keskmine lemmapikkus ning tajuverbide esinemine.

Väga spontaanne (inimeste hinnang) vs vähe spontaanne (automaatne hindamine)

Ei leidunud ühtegi teksti, mis oleks inimeste poolt hinnatud spontaanseks, aga automaatselt hinnatud mittespontaanseks. Siiski leidus tekste, mis olid automaatselt hinnatud vähem spontaanseks, kuid inimeste poolt väga spontaanseks. Selline tekst on näide 7, mis on võetud Lapsemure foorumist. Selle inimhinnatud skoor on 0.95 ja automaatselt hinnatud skoor 0.51:

(7) *Mul on selline mure ,et mul on üks klassiõde alguses tundus ta normaalne olime parimad sõbrad olime kogu aeg koos.2klassis juhtus selline asi ,et ta varastas minu poolt ühe asja pärast kui mu ema seda ta ema käest tagasi küsis ei uskunud ta ema seda ja mu klassiõde rääkis mu emaga päris ebaviisakalt . Pärast seda hakkas klassiõde minuga ülbitsema . Pärast läks asi veel hullemaks ja ta hakkas mind solvama ,aias mu sõbrad minu juurest ära(ma sain enne kõigiga hästi läbi) ja siis hakkasid need teised klassiõed mind ka juba solvama ja ta rääkis igasuguseid asju minu kohta.Lõpuks kui me hakkasime eraldi vahetama (kolmanda klassi algul)poistega(algul tüdrukud ja siis poisid)hakkasid nad tütrukute vahetuse ajal mind igat moodi solvama,minu ees ülbitsema.Kui ma ütlesin ,et räägin õpetajale ütlesid nad ,et kes mind usub kui nemad väidavad vastu pidist.Nii see käib siiamaani kuigi mul on nüüd kaks sõpra(tüdrukud) üks teine põlaalune ja mu pinginaaber.Nad on normaalsed.Vahepeal proovisin ka selle klassiõe meele järgi olla ,aga*

see ei tulnud mul hästi välja ja ta oli mind juba tõsiselt vihkama hakkand. Palun aidake!Ma ei tea mida teha! Tänudega Katri! (etTenTen www.lapsemure.ee, doc id = 140896)

Tekstis on 13 kokkukleepunud kirjavahemärki, 1,35% sõnadest on kirjavigadega, 67,6% asesõnadest väljendavad esimest isikut ning keskmine lemmapikkus on 4.23 tähte. Tegu on nende tunnuste poolest väga spontaanse tekstiga. Kuid tekstis ei esine korduvaid tähti ega suuremaid sõnaüksuseid, puuduva suure algustähega või läbinisti suurte tähtedega sõnu. Nende tunnuste puudumise tõttu on automaatne hinnang palju madalam, kui inimhinnang.

4.2. Järeldused ja võimalikud edasiarendused

Selles töös kirjeldasin esimest katset määrata automaatselt tekstide formaalsust ja spontaansust. Formaalsuse ja spontaansuse automaatse hindamise süsteem andis palju informatsiooni, kuidas kasutatud tunnused ja nende osakaal lõplikust hinnangust seostuvad inimhinnangutega. Seda informatsiooni saab tulevikus kasutada automaatse hindamise edasiarendamiseks, et ümber hinnata kasutatavate tunnuste osakaalu ja määrata kasutatud tunnuste komplekti piisavust või ebapiisavust.

Praegusel hindamisel on kaks peamist probleemi. Esimene probleem on hinnangute nulli lähedale jäämine. Seda probleemi saab lahendada mitmel viisil, näiteks määrates maksimaalse ja minimaalse piiri, kui arvutada punktisumma ümber hinnanguks. See oleks kasulik, kuna väga formaalsetes ja spontaansetes tekstides ei pruugi olla kõiki positiivseid tunnuseid või ei pruugi neid olla piisavalt, et anda maksimaalne punktisumma: et teksti spontaansuse hinnang oleks 1, peaks olema tekstis kirjavigadega sõnu, kokkukleepunud kirjavahemärke, läbivalt suurte tähtedega ning puuduva suure algustähega sõnu, emotikone ja tajuverbe, kuid inimhinnangu järgi annab maksimaalse spontaansuse skoori ka ainult mõne tunnuse rohke esinemine nende hulgast.

Alternatiivne võimalus esimese probleemi lahendamiseks oleks ühe- ja kahesuunaliste tunnuste piiride nihutamine, et null punkti andev vahemik oleks kõige väiksem ja viis punkti andev vahemik oleks kõige suurem, vahepealsete gruppide vahemikud kasvaksid vastavalt. See suurendaks võimalust, et tekst saab nullist erinevama hinnangu. Kolmas võimalus oleks, et iga tunnus mõjutaks skoore eri määradel, ehk et mõned tunnused oleksid hindamisel mõjukamad, kui teised.

Piiride nihutamise, punktide mõjukuse varieerimine ja punktisummale maksimaalse ja minimaalse piiri lisamise võimalusi saab ka koos rakendada.

Teine probleem esineb vaid spontaansuse hindamisel, kuna sellel on palju rohkem positiivselt mõjutavaid tunnuseid kui negatiivselt mõjutavaid. Probleem väljendub tekstidel, milles esinevad üksikud spontaansuse tunnused, nagu vähene sõnavara varieerumine või rohke esimese isiku asesõnade kasutamine, kuid negatiivse mõjuga tunnused ei tühista valesti määratud spontaansust ja seega hinnatakse teksti liiga spontaanseks. Seda probleemi on kõige mõistlikum lahendada lisades hindamisele juurde tunnuseid, nagu sõnavara ja süntaksi analüüs, mis aga sellest tööst jäid välja. Alternatiiv oleks anda negatiivsete tunnuste eest rohkem punkte, et nendel rohkem mõju oleks, kuid sellega on oht, et ühe negatiivse tunnusega tühistatakse mitme positiivse tunnuse mõjud, näiteks kõrge sõnavara variatiivsus tühistaks suure kirjavigade, emotikonide ja korduste arvu.

Automaatse hindamise edasiarendamiseks on mitmeid võimalusi. Üks võimalus oleks masinõppe kasutamine. Sheikha ja Inkpen (2012) on varasemalt uurinud, kuidas formaalsust hinnata leksikaalgrammatiliste tunnuste abil kasutades masinõpet ning millised tunnused olid selleks kõige kasulikumad. Et aga analoogselt masinõppe abil eesti keeles formaalsust ja spontaansust hinnata, oleks vaja testkorpusest suuremat treeningandmestikku, kus oleksid inimesed hinnanud tekstide määratud formaalsust ja spontaansust. Selle treeningandemstiku jaoks on vaja aga teha eraldi uurimus.

5. Kokkuvõte

Töö eesmärk oli luua automaatne hindamissüsteem, mis mõõdaks Interneti tekstide spontaansust ja formaalsust kui Vaigu jt (2020) dimensionaalse tekstimudeli dimensioone. Selle jaoks otsisin esmalt kirjandusest spontaansust ja formaalsust kirjeldavaid tunnuseid ning valisin nendest alamhulga, mida automaatsel hindamisel tunnuste komplektina kasutada. Lõin automaatse hindamise süsteemi, mis hindab sisendtekstide spontaansust ja formaalsust skaalal -1–1. Süsteemi evalveerisin kasutades inimhinnangutega korpust. Võrdlesin automaatset hinnangut inimhinnanguga ning uurisin hinnangute erinevuste põhjuseid.

Töö praktiline eesmärk on täidetud, loodud on kasutatav süsteem, mis annab igale sisendtekstile spontaansuse ja formaalsuse hinnangu. Töö teoreetilised eesmärgid said samuti täidetud. Töös defineerisin mitmeid tunnuseid, mis kirjeldavad tekstide spontaansust ja formaalsust. Tunnus võis teksti kirjeldada nii positiivselt, väljendades formaalsust või spontaansust, kui ka negatiivselt, väljendades mitteformaalsust või mittespontaansust. Nendest koostasin kaks komplekti, ühe formaalsuse hindamiseks ja teise spontaansuse hindamiseks.

Formaalsuse hindamise tunnused olid sõnavara mitmekesisus, keskmine lemma ja liitsõna osalemma pikkus, esimese ja teise isiku esinemine, kolmanda isiku esinemine, emotikonid, kaudne kõneviis, *nud*-partitsiip, impersonaal ning nimisõnafraaside pikkus ja arv. Spontaansuse hindamise tunnused olid sõnavara mitmekesisus, keskmine lemma ja liitsõna osalemma pikkus, esimese isiku asesõnade kasutamine, emotikonid, kirjavead, tajuverbid, tühikuta kirjavahemärgid, läbivalt suurtähtedega kirjutatud sõnad ja puuduva väikese algustähega sõnad.

Tunnuste komplekte rakendasin loodud automaatses hindamissüsteemis, saades igale tekstile nii formaalsuse kui ka spontaansuse hinnangu. Seejärel võrdlesin automaatset hindamist inimeste hinnatud testkorpusega. Inimeste hinnangutega võrreldes sain teada, kuidas valitud tunnuste komplekti kasutades erines ja sarnanes automaatne hindamine. Erinevused inimhinnangu ja automaatse hindamise vahel on suures osas põhjustatud sõnavara ja lausestruktuuri kui tunnuste mitte vaatamisest.

Töös esitatud katse analüüsi tulemusel saab täpsemalt planeerida edasisi töösuundi ning sain teada, et kuigi kasutatud tunnuste komplekt sisaldab õigeid tunnuseid, pole see piisav ja vajab seega leksikaalsete ja süntaktiliste tunnuste lisamist ning peenhäälestamist. Sõnavara ja lausestruktuur osutusid formaalsuse ja spontaansuse hindamisel väga olulisteks tunnusteks, mida ei olnud komplekti valitud. Automaatse hindamise edasiarendamisel tasuks tuua need tunnused sisse. Hindamist saaks ka edasi arendada, katsetades dimensioonide skooridele piiride lisamist ning varieerides tunnuste mõju dimensiooni skoorile.

Lisaks on idee luua eraldi testkorpus, kus oleksid tekstid inimeste formaalsuse ja spontaansuse hinnangutega. See oleks kasulik, et saada rohkem andmeid automaatse hindamise võrdlemiseks inimeste hinnangutega või et oleks võimalik kasutada suurt korpust masinõppeks. Mõlema ülesande jaoks oleks vaja aga suurt treeningandmestikku, mis tuleks aga nende jaoks eraldi luua.

6. Kasutatud allikad

6.1. Kirjandus

Kerge, Krista, Hille Pajupuu 2010. Text-types in speech technology and language teaching. – Analizar datos > Describir variación / Analysing data > Describing variation. Toim Jorge L. Bueno Alonso jt. Vigo: Universidade de Vigo, Servizo de Publicacións, 380−390.

Laippala, Veronika, Jesse Egbert, Douglas Biber, Aki-Juhani Kyröläinen 2021. Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. – Lang Resources & Evaluation. https://link.springer.com/article/10.1007/s10579-020-09519-z

Laur, Sven, Siim Orasmaa, Dage Särg, Paul Tammo 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline. – Proceedings of The 12th Language Resources and Evaluation Conference. Marseille: European Language Resources Association 7154–7162.

Lindström, Liina, Piret Toomet 2000. Eesti suuliste narratiivide keelelisi erijooni. – Eesti keele allkeeled. Toim. Tiit Hennoste. Tartu: Tartu Ülikooli Kirjastus, 174−203.

Muischnek, Kadri, Heiki-Jaan Kaalep, Raul Sirel 2011. Korpuslingvistiline lähenemine eesti internetikeele automaatsele morfoloogilisele analüüsile. – Eesti Rakenduslingvistika Ühingu aastaraamat 7, 111−127.

Reinsalu, Riina 2011. Lepingute lausestruktuur. – Emakeele Seltsi aastaraamat 57, 218−234.

Santini, Marina, Alexander Mehler, Serge Sharoff 2011. Riding the Rough Waves of Genre on the Web. – Genres on the Web. (Text, Speech and Language Technology 42.) Toim Alexander Mehler, Serge Sharoff, Marina Santini. Dordrecht: Springer Publishing Company, 3–30.

Sharoff, Serge 2021. Genre Annotation for the Web: Text-external and text-internal perspectives. – Register studies 3, 1–32.

Sheikha, Fadi Abu, Diana Inkpen 2012. Learning to Classify Documents According to Formal and Informal Style. – Linguistic Issues in Language Technology 8/1, 1−29.

Vaik, Kristiina, Kairit Sirts, Kadri Muischnek 2020. Dimensionaalne tekstimudel. Teoreetiline ülevaade. – Keel ja Kirjandus 10, 875−898.

6.2. Veebivarad

Eesti WordNet. https://doi.org/10.15155/1-00-0000-0000-0000-0013AL

EstNLTK. https://github.com/estnltk/estnltk

etTenTen. https://doi.org/10.15155/1-00-0000-0000-0000-0011FL

Looks.wtf. https://looks.wtf/. Vaadatud 06.04.2021.

The Unicode Consortium 2021. Unicode® Emoji Charts v13.1. Mountain View, CA: The Unicode Consortium. http://www.unicode.org/emoji/charts-13.1/

Wikipedia. https://en.wikipedia.org/wiki/List_of_emoticons. Vaadatud 05.04.2021.

Ühendkorpus 2019. https://doi.org/10.15155/3-00-0000-0000-0000-08565L

Ühendsõnastik 2021. Koostanud ja toimetanud Indrek Hein, Jelena Kallas, Olga Kiisla, Kristina Koppel, Margit Langemets, Tiina Leemets, Maia Melts, Sirje Mäearu, Tiina Paet, Peeter Päll, Maire Raadik, Mai Tiits, Katrin Tsepelina, Maria Tuulik, Udo Uibo, Tiia Valdre, Ülle Viks. Eesti Keele Instituut. https://doi.org/10.15155/3-00-0000-0000-0000-08979L

7. Summary. Evaluating the spontaneity and formality of online texts as dimensions of the dimensional text model.

The goal of this thesis was to experiment with automatic evaluation of the spontaneity and formality of online texts as dimensions of the dimensional text model created by Vaik et al. (2020). The practical goal of this experiment was a functional spontaneity and formality evaluation system which brought forth two theoretical goals: creating a list of features describing the dimensions using preexisting research and selecting a subset of the features to use for evaluation.

A functional evaluation system for spontaneity and formality was created. For both dimensions, formality and spontaneity, a selection of features was made based on previous research. The features can influence the evaluation both positively or negatively and the score given by the system will be between negative one and one, where negative one represents anti-spontaneity or anti-formality and one represents formality.

The system was evaluated on a test corpus where the formality and spontaneity of the texts had been evaluated by humans. As a result of the experiment and its evaluation it was possible to discover that although the selection of features used in this experiment included accurate features, the selection is not sufficient and requires additional lexical and syntactic features as well as fine-tuning. Additionally, in order to develop the evaluation system further there is a need to create a new corpus where the texts have been evaluated on their formality and spontaneity by humans.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Karl Gustav Gailit

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose
„SPONTAANSUSE JA FORMAALSUSE KUI DIMENSIONAALSE TEKSTIMUDELI DIMENSIOONIDE AUTOMAATNE HINDAMINE VEEBITEKSTIDES“,

mille juhendajad on Kadri Muischnek ja Kristiina Vaik

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.

Karl Gustav Gailit

20.06.2021