Tartu Ülikool

Humanitaarteaduste ja kunstide valdkond

Eesti ja üldkeeleteaduse instituut

Karl Gustav Gailit

LEKSIKONIDE JA KAALUDE LISAMINE VEEBITEKSTIDE FORMAALSUSE JA SPONTAANSUSE DIMENSIOONIDE HINDAMISE MUDELI ARENDAMISEKS

Magistritöö

Juhendajad Kristiina Vaik ja Kadri Muischnek

Tartu 2023

Sisukord

Autorsuse kinnitus

Kinnitan, et olen käesoleva lõputöö ise kirjutanud ning toonud korrektselt välja teiste autorite panuse. Töö on kirjutatud lähtudes Tartu Ülikooli eesti ja üldkeeleteaduse instituudi lõputöö nõuetest ning on kooskõlas heade akadeemiliste tavadega.

Sissejuhatus

On vaieldamatu tõde, et nüüdisaegne maailm leiab aset suures osas Internetis. Paljud vestlused toimuvad sõnumivahetustena või meilidena, reisipäevikud laetakse üles blogidena, teadusartiklid laetakse üles kõigile nägemiseks ning kõikidel lehekülgedel on oma kasutustingimused. Ka keeleteadus toimub Internetis, eriti korpuslingvistika, kus vanu korpuseid laetakse üles, et neid säilitada ja ligipääsetavamaks teha. Internetis leiduvatest tekstidest moodustatakse ka pidevalt uusi korpuseid, et saada suures koguses tekste, et neid kasutada kvantitatiivseks analüüsimiseks või masinõppeliste mudelite treenimiseks.

Kuid sellisel viisil korpuste loomisel jääb mitmeteks uurimusteks informatsiooni puudu. Üks selline puudujääv osa informatsioonist on korpuses olevate tekstide liigid. Varasemad korpused olid reeglina žanripõhised, sisaldades näiteks vaid ajakirjandust, suuliste murdekeeleliste vestluste transkriptsioone või ilukirjandust, mis on avaldatud 1930. aastatel. Veebitekste kokku kraapides on aga sarnaseim informatsioonikild teksti allikaks oleva lehekülje URL. Kuigi esmapilgul tundub, et ajakirjanduslikke tekste oleks võimalik välja noppida vaadates, kas tekst on pärit näiteks Postimehe või Eesti Rahvusringhäälingu lehekülgedelt, ei ole see siiski lahenduseks. Üks probleem on, et ei ole teada, kas see tekst on artikkel ise või sellega seotud kommentaarid. Lisaks kasutatakse näiteks blogihaldusplatformi Blogspot nime lisaks blogidele ka mitmetel teistel eesmärkidel, sealhulgas ilukirjanduse avaldamiseks ning vanade murdetekstide arhiveerimiseks.

Seepärast on veebitekstide liigitamine oluline ja endiselt aktuaalne probleem, mille lahendamiseks on pakutud mitmeid meetodeid. Üheks selliseks meetodiks on Kristiina Vaik jt (2021) välja pakkunud dimensionaalse tekstimudeli, mis klassifitseerib tekste nende dimensioonide alusel. Nende dimensioonide hulka kuuluvad formaalsus ja spontaansus, mille hindamiseks tegin oma bakalaureusetöös (Gailit 2021) mudeli. Loodud mudel oli toimiv, andes adekvaatseid tulemusi, kuid soovisin mudelit edasi arendada, sest mitmed olulised tunnused ja aspektid olid mudelist välja jäänud.

Selles töös lisan ma mudelile kaks aspekti: sõnavara vaatlemise leksikonide alusel ja tunnuste olulisuste varieerimise kaalude lisamise kaudu. Nende kahe aspekti mudelile lisamine oli väga ajamahukas, mistõttu oleks ülejäänud kümne dimensionaalse tekstimudeli dimensioonide hindamiseks kasulik teada, kas leksikonid ja kaalud on kasulikud ja tarvilikud, ning need vastavalt kas kasutusele võtta või välja jätta.

Magistritöö eesmärk on edasi arendada varasemalt loodud mudelit, uurida selle tulemusi ning analüüsida, kas ja kuidas on edasiarendused mõjutanud mudeli tulemusi. Mudeli analüüsi ja selgituse kõrval leidub ka kasutatav programm ise, mida on võimalik arvutisse alla laadida GitHubi[1] kaudu.

Magistritöö koosneb viiest osast. Töö esimene osa *Tekstiliikide klassifitseerimine* tutvustab varasemaid tekstide liikide alusel klassifitseerimise mudeleid ning keskendub pikemalt dimensionaalsele tekstimudelile (Vaik jt 2020), mille formaalsuse ja spontaansuse dimensioonide hindamiseks töö mudel on loodud. Teine osa *Andmestikud* annab ülevaate töös kasutatud andmestikest. Kolmas osa *Tunnused* kirjeldab mudelis kasutatavaid tunnuseid. Kirjeldan pikemalt spontaansust ja mitteformaalsust väljendavate sõnade leksikonide loomist ja tunnusena kasutamist. Neljas osa *Tunnuste kaalud* kirjeldab mudelis kasutatud tunnuste kaalude määramise protsessi ning millised kaalud lõplikus mudelis on kasutatud. Viies ja viimane osa *Tulemused* kirjeldab, kuidas sisse viidud muutused on mõjutanud mudeli hinnanguid, võrreldes bakalaureusetöö mudeli hinnangutega.

[1] Link projekti GitHubile: https://github.com/kgailit/Gailit-2023

1. Tekstiliikide klassifitseerimine

Tekstiliigid on korpuslingvistikas väga oluline informatsioon, sest need näitavad, mis tüüpi tekste vaadeldav korpus sisaldab, ning seekaudu võimaldavad kasutajatel otsustada, kas tegu on nende jaoks kasuliku korpusega. Sellised on vanemad, teise põlvkonna korpused, mis on käsitsi valitud ning neid koostades on igale tekstile pandud märge, mis liiki tekstiga on tegu. Üks näide teise põlvkonna korpusest on eesti keele koondkorpus[2] ja selle allosa tasakaalus korpus[3], mis koosneb võrdses mahus kolmest tähtsaimast kirjaliku keele tekstiklassist: ajalehetekstidest, ilukirjandusest ja teadustekstidest. Kolmanda põlvkonna korpused, mis koosnevad Internetist korjatud tekstidest, on teise põlvkonna korpustest palju suuremad ning on seega kasulikud, kui andmete kvantitatiivsus on vajalik. Kuna need korpused on aga automaatselt kogutud, ei ole nendes inimeste koostatud korpustega võrreldes sama palju metainformatsiooni, sealhulgas tekstide liike. (Muischnek, 2015: 38)

Tekstiliikide informatsioon on aga kasulik informatsioon paljude ülesannete jaoks, nagu kindla tekstiliigi keelekasutuse uurimine või analüüsimine, mille alusel mõni tunnus varieerub kirjalikus keelekasutuses, mille tõttu on tegu aktuaalse probleemiga. Kuid veebitekstide tekstiliikide klassifitseerimise probleem koosneb mitmest osast: veebitekstide taksonoomia, tekstide hübriidsus ning klassifitseerimine kui protsess ise.

Tekstide taksonoomia probleem leidub ka trükitud tekstide puhul, kus tekstide liigitamine toimub erinevalt olenevalt teadusalast ja tekstiuurijast, mistõttu ei leidu ühte definitiivset loetelu kõikidest tekstiliikidest ega žanritest. Siiski leidub aga väga üldine viis eristada keelekasutusvaldkondi: argikeel, ilukirjanduskeel ja tarbekeel. Need kasutusvaldkonnad on omakorda oma alaliikidega, mis ise omavad mitmeid alaliike, näiteks tarbekeele alla kuuluvad teaduskeel ja ajakirjanduskeel ning ajakirjanduskeele enda alla kuuluvad näiteks uudised, intervjuud ja arvamuslood. (Kasik 2007: 35)

[2] Koondkorpus: https://www.cl.ut.ee/korpused/segakorpus/
[3] Tasakaalus korpus: https://www.cl.ut.ee/korpused/grammatikakorpus/

Veebitekstid võivad kuuluda kõikidesse eelmainitud keelekasutusvaldkondadesse, kuid nende puhul leidub ka mitmeid teisi liike, mis paberil ei ole võimalikud. Sellised on näiteks mitme autori koostööl suhtlemise eesmärgil kirjutatud lühikestest tekstijuppidest koosnevad tekstid, nagu foorumipostitused, jututoad ja muud veebisuhtluse vahendid ning sotsiaalmeedia. Veebitekstide taksonoomia probleemi saab täheldada ka nende tekstiliikide klassifitseerimise mudelite puhul, kus mudelitel on klassifitseeritavate klasside arv varieeruv, näiteks Veronika Laippala jt (2021) mudel kasutab 26 registrit ning Noushin Rezapour Asheghi jt (2014: 41) kasutab 15 žanrit.

Taksonoomia kõrval esineb veebitekstidel teinegi aspekt, mis teeb nende liikide määramise keerukamaks – tekstiliikide hübriidsus (vt nt Santini 2010, Laippala jt 2022). See tähendab, et kindel üksiktekst võib kuuluda mitme tekstiliigi alla, näiteks kuuluvad blogipostituste ning juhendite, täpsemalt retseptide, liikide alla retseptiblogides leiduvad tekstid, mis sisaldavad mõnda retsepti kohta käivat lugu autori elust, retsepti jutustavat kirjeldust koos näpunäidetega ning retsepti kui täpsete arvude ning sammudega juhendit ennast.

Veebitekstide tekstiliikide klassifitseerimise kolmas lahendamata osa on klassifitseerimise protsess ise. Kui vanemad korpused on koostatud ja seega liigitatud käsitsi, on kolmanda põlvkonna korpused massilised, mistõttu ei ole see variant internetikorpuste jaoks mõttekas. Seetõttu on probleemi lahendamiseks välja pakutud mitmeid teisi meetodeid, millest valdav enamus on masinõppelised mudelid. Pakutud mudelid põhinevad peamiselt n-grammidel ehk tekstis kõrvuti asetsevatel tähemärkidel, lihtsustatult tekstis kasutatud sõnavormidel. See aga tähendab omakorda, et tekste liigitatakse nende sisude järgi ning mudelid kipuvad hindama treeningkorpuste väliseid tekste palju kehvemini. (Laippala jt 2021: 758–762, Sharoff 2021: 2–4)

Ülesobitamise probleemi leevendamiseks on *n*-grammidel põhinevaid mudeleid edasi arendatud. Andmestiku suurendamine ja varieerimine aitab probleemi leevendada, kuid lisainformatsiooni, nagu morfosüntaktilise informatsiooni, kasutamine annab veelgi paremaid tulemusi. Seda tegid näiteks Laippala jt (2021), kus 26 registri liigitamise

mudel, mis kasutas nii leksikaalset kui ka grammatilist informatsiooni, sai keskmiseks F1-skooriks 74,51%.

Rezapour Asheghi jt (2014) võrdlesid kahte treeningandmestikku, üht ainult tähemärkide *n*-grammidega ja teist, mis sisaldas ka lisainformatsiooni, nagu informatsiooni teksti grammatika kohta ja statistikat, sealhulgas sõnavara variatiivsus, nimeüksuste sagedus, keskmine sõna- ja lausepikkus, keskmine silpide arv sõnas ning HTML märgendite sagedus. Treenides kummagi andmestiku jaoks tekstide klassifitseerimise mudelid, sai ainult *n*-grammidest koosnev mudel täpsuseks 78,88% ning lisainformatsiooniga mudel sai 89,63%. Sellest saab järeldada, et lisainformatsioon on väga oluline heade tekstide liigitamise mudelite treenimiseks.

Veebitekstide liigitamise probleemi lahendamiseks on Vaik jt (2020) välja pakkunud dimensionaalse tekstimudeli, mis jagab tekstid keeleliste tunnuste komplektide alusel mõõdetud tuumiknähtuste ehk dimensioonide kimpudeks. Tekstid, mille koosesinevate dimensioonide kimbud on sarnased, on ka funktsioonide poolest sarnased ning seega võiksid need kuuluda ühte tekstiliiki. Kuna mudel põhineb sisendteksti alusel arvutatud tunnustel ja mitte sisendtekstil endal, ei sõltu teksti hinnang mudeli treenimiseks kasutatud andmestikust. See tagab ka, et teksti liik ei sõltu vaid sisust, mis võimaldab stabiilsemat ja seega täpsemat klassifitseerimist.

Dimensionaalses tekstimudelis on kaksteist iseseisvat, üksteisest sõltumatut dimensiooni: abstraktsus, afektiivsus, instrueerivus, informatsioonitihedus, spontaansus, formaalsus, impersonaalsus, ajalisuse olulisus, interaktiivsus, subjektiivsus, keerukus ja argumentatiivsus. Kuigi dimensioonid on iseseisvad, võivad nende tunnused omavahel osaliselt kattuda. (Vaik jt 2020: 882–890)

Selles töös olen edasi arendanud oma bakalaureusetöös (Gailit 2021) loodud programmi, mis hindab tekstide formaalsuse ja spontaansuse dimensioone. Edasi annan lühikese ülevaate mõlemast dimensioonist.

1.1. Formaalsus ja spontaansus

Formaalsus kui dimensioon on iseloomulik ametliku keelekasutusega tekstidele. Dimensionaalse tekstimudeli kirjelduses on Vaik jt (2020: 887) oletanud, et tekstides väljendatakse formaalsust peamiselt leksikaalselt viisakusväljendite kasutuse ja kõnekeelsuse vältimise kaudu. Lisaks kasutatakse keeruka ehitusega lauseid ning rohkelt nominalisatsioone. Tekstid, milles esineb palju formaalsust, on näiteks lepingud ja protokollid, teadusartiklid ning ka näiteks viisakad meilid ja ametlikud kõned, ning formaalsust ei esine kõnekeelsetes tekstides, nagu blogipostitustes, harilikes jututubade vestlustes ja sõnumivahetustes.

Tekstide formaalsuse automaatset hindamist on varasemalt uurinud Francis Heylighen ja Jean-Marc Dewaele (1999), pakkudes välja *F*-skoori kui teksti formaalsuse empiirilise hinnangu. Nende mudel põhines ainult sõnaliikide infol, kus, nagu nominaalstiilile on omane, väljendavad nimisõnad, omadussõnad, artiklid ja kaassõnad formaalsust ning asesõnad, adverbid ja hüüdsõnad ebaformaalsust. *F*-skoori kui formaalsuse käsitlus sarnaneb selles töös kasutatava käsitlusega, kus formaalsus väljendab täpsust ja skaala äärmises otsas ka kindlate väljendusvormelite kordamist, nagu seadustekstides samale olukorrale või nähtusele viitamisel kasutatakse alati samu sõnu või fraase. Tegu ei ole aga kattuvate formaalsuse käsitlustega, kuna *F*-skoor on seotud ka keerukuse ja informatsioonitihedusega, mis on dimensionaalses tekstimudelis iseseisvate dimensioonidena, mitte formaalsuse osadena. Formaalsust on uurinud ka Fadi Abu Sheikha ja Diana Inkpen (2012), kuid kuigi nendelgi on formaalsus tugevalt seotud keerukusega, on nad käsitlenud tunnustena rohkemaid tunnuseid, mida olen lisanud ka enda mudelisse. Mudelile formaalsuse tunnuseid otsides olen uurinud ka stereotüüpsete formaalsete tekstide kirjeldusi, nagu Riina Reinsalu (2011) lepingute lausestruktuuri.

Spontaansus kui dimensioon on iseloomulik tekstidele, mis on esitatud vahetult ja toimetamata. Vaik jt (2020: 886) on oletanud, et tekstides väljendatakse spontaansust peamiselt lihtsa ning mittestandardse keelekasutusega, sealhulgas vigaderohkus aga ka tahtlikud asendused ja lühendamised. Palju spontaansust sisaldavad tekstid on näiteks jututubade vestlused ning sõnumivahetused, kuna nende kirjutamisel on olnud reaalajalisi

piiranguid. Spontaansust kui dimensiooni ei leidu aga toimetatud tekstides, nagu teadusartiklites ja lepingutes, ega ka ilukirjanduses.

Erinevalt formaalsusest, ei ole spontaansuse hindamiseks teadaolevalt varasemalt loodud automaatseid mudeleid. Seda on aga tehtud ebaformaalsusega (inglise keeles *informality*), kus Mosquera ja Moreda (2011) on uurinud Web 2.0 tekste, täpsemalt sotsiaalmeedia postitusi, blogisid ja suhtluskeskkondi, ning nendele omaseid tunnuseid. Need tekstiliigid kuuluvad stereotüüpsete tugevalt spontaansete tekstide hulka ning mitmeid nende mainitud tunnuseid, nagu kõnekeelsus ja emotikonide kasutus, olen samuti käsitlenud kui spontaansuse tunnuseid.

Spontaansust väljendavate tunnuste leidmiseks mudelis kasutamiseks olen vaadelnud artikleid, mis käsitlevad stereotüüpseid spontaanseid tekstiliike. Selline on näiteks Kadri Muischneki jt (2011) uurimus internetikeele morfoloogilisest analüüsimisest, mis kirjeldab veebitekstide, nagu jututubade vestluste, tüüpilisi omadusi. Lisaks on kõikide tekstide hulgast spontaansed tekstid kõige sarnasemad suulisele kõnele, olles mõjutatud toimetamata olekust ja kirjutamisel esinenud ajapiirangutest. Seepärast olen vaadelnud ka võrdlusi suulise kõne ja tekstide vahel, nagu Lindströmi ja Toometi (2000) kirjeldust suuliste ja kirjalike narratiivide eripäradest.

Kuigi formaalsuse ja spontaansuse dimensioonid võivad tunduda kui ühe skaala kaks eri otsa, ei ole see dimensionaalse tekstimudeli puhul tõene. Paljude tekstide puhul kehtib olukord, et mida kõrgem on teksti formaalsus, seda madalam on spontaansus, ja vastupidi. Sellised tekstid on näiteks eelmainitud teadusartiklid ja lepingud, kus formaalsus on tugev ja spontaansus olematu, ning jututubade vestlused ja blogipostitused, kus spontaansus on tugev ja formaalsus olematu. On aga ka mitmeid tekstitüüpe, kus see olukord ei kehti. Näiteks on viisakad meilid tugevalt formaalsed, kuna nendes kasutatakse viisakusvormeleid ja pöördumisi, kuid inimestevahelise suhtlemise vahendina on meilid sageli kirjutatud kiiresti ja seega spontaansuserohkelt. Ilukirjandus, kus tekstid on põhjalikult toimetatud, olles üle vaadatud mitte ainult autori enda aga ka keeletoimetajate poolt, ei sisalda reeglina spontaansust ning need väljendavad ka formaalsust kas nõrgalt või üldsegi mitte, olles loodud kergesti loetavaks meelelahutuseks.

Kuna eri tekstiliikide puhul esinevad dimensioonid erinevalt, olen üritanud kasutada andmestikke, mis sisaldaksid tekste võimalikult erinevatest liikidest.

2. Andmestik

Mudeli koostamiseks olen kasutanud kolme korpust.

Esiteks kasutan Ühendkorpuse[4] 2019. aasta veebitekstide alamkorpust, millest olen võtnud 100 000 teksti suuruse alamhulga ja mida edaspidiselt nimetan Ühendkorpuse valimiks. Valim koosneb Ühendkorpuse eeltöötlemata variandist, kus tekstid on vaid lausestatud ja on seega võimalikult sarnased otse veebist kogutud tekstidega. Alamhulka olen tekstid valinud suvaliselt, kuid püüdsin kindlustada võimalikult suurt allikate varieeruvust, valides esmalt suvalise allika ja seejärel ühe suvalise teksti, mis allikas leidus. Jätsin välja kõik tekstid, mis olid alla 500 tähemärgi, sealhulgas erisümbolid, nagu tühikud ja kirjavahemärgid. Lisaks lühendasin tekste, mis olid üle 200 lause pikad, jättes tekstiks vaid esimesed 200 lauset. (Gailit 2021: 9)

Teiseks kasutan oma juhendajalt, Kristiina Vaigult, saadud inimhinnangutega korpust, mis koosneb 120 tekstist. Need tekstid pärinevad etTenTen13[5] korpusest. Korpuses on kõik tekstid annoteeritud dimensionaalse tekstimudeli kõikide dimensioonide skooridega, seega on tegu andmestikuga, mis sisaldab formaalsuse ja spontaansuse hinnanguid, mis on määratud samadel põhimõtetel, nagu mudeli loomiseks on vaja, ning mis seega on sobilik mudeli arendamiseks ja hindamiseks.

Korpust hinnates ehk annoteerides pidid annoteerijad hindama tekste dimensioonide kolmikute kaupa. Neile esitati üheaegselt kaks teksti (tekst A ja tekst B), millest tuli valida vaadeldavale dimensioonile kõige iseloomulikum tekst ja hinnata 4-pallisel järjestusskaalal: 0 ehk dimensioon tekstis puudub; 1 ehk dimensiooni esineb nõrgalt; 2 ehk dimensiooni esineb mõõdukalt, ning 3 ehk dimensiooni esineb tekstis tugevalt. Tekst, mida ei valitud, sai valitud tekstist ühe võrra madalama hinnangu. Kui annoteerija ei

[4] Ühendkorpus: https://doi.org/10.15155/3-00-0000-0000-0000-08565L
[5] etTenTen13: https://doi.org/10.15155/1-00-0000-0000-0000-0011FL

suutnud tekstide vahel valida, said mõlemad tekstid hinnanguks nulli. Teksti lõplikuks hinnanguks on sellele antud annoteeringute aritmeetiline keskmine, kuid analüüsimise lihtsustamiseks olen need jaganud kolmega, et vaadeldav vahemik oleks null kuni üks, kus null on dimensiooni puudumine ja üks dimensiooni tugev esinemine.

Annoteerimise ülesehituse tõttu on korpuses eri tekstidel eri arv hinnanguid. On tekste, mis said kõigilt annoteerijatelt nullist suurema hinnangu, aga on ka tekste, mida ei ole ükski annoteerija kordki valinud ning seega on saanud hinnanguks nulli. Tuleb aga täheldada, et kui teksti keskmine hinnang on null, võib sellel olla teisi põhjuseid kui see, et dimensiooni tekstis ei leidugi. Näiteks võib tekstijupp olla annoteerijale segaseks jäänud või ei ole kahe võrdse teksti vahel suudetud otsustada.

Kolmandaks andmestikuks on 60 teksti eelnevalt kirjeldatud Ühendkorpuse valimist, millele olen ise andnud formaalsuse ja spontaansuse hinnangu vahemikus 0 kuni 1. Tekstid said valitud bakalaureusetöö mudeli skooride alusel: 10 kõrge skooriga teksti, 10 madala skooriga teksti ja 10 keskmise skooriga teksti, nii formaalsuse kui ka spontaansuse skooride alusel. Tekstid on hinnatud ainult ühe inimese poolt, seega nende hinnangud ei ole nii täpsed kui mitme inimese hinnangute keskmised, kuid nende lisamine oli vajalik, sest need tekstid sisaldasid mitmeid tunnuseid, mis inimhinnangutega korpuses esinesid vähesel määral.

Viimast kahte korpust kasutan koos, kuna mõlema ülesanne on väljendada formaalsuse ja spontaansuse inimhinnanguid, mida magistritöös loodav mudel peab ennustama. Nimetan seda kooskasutust hinnangukorpuseks.

3. Tunnused

Mudeli alustalaks on varasemad tekstiliikide uurimisel tuvastatud liikidele iseloomulikud keelelised ja tekstilised tunnused ning nende arvulised representatsioonid. Iga tunnus on kasutatud kas ainult formaalsuse, ainult spontaansuse või mõlema dimensiooni skoori arvutamisel. Edasi loetlen ja selgitan kõiki tunnuseid, mida mudelis kasutan, ja toon välja,

kuidas need mõjutavad hinnangut. Kõik tunnused, peale leksikonide, on juba eelnevalt kasutatud ja kirjeldatud minu bakalaureusetöös (Gailit 2021).

3.1. Varasemad tunnused

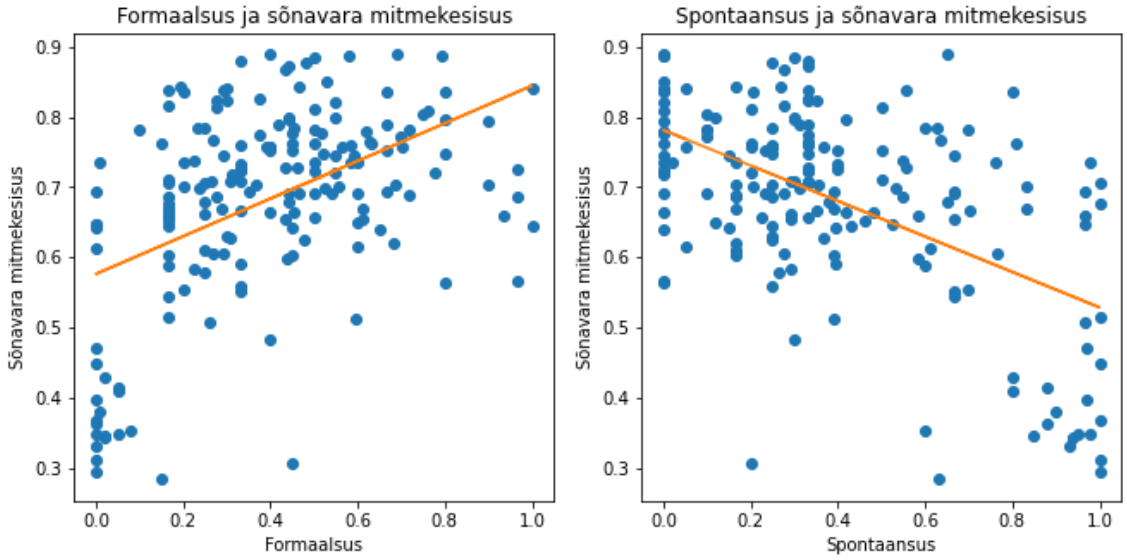
3.1.1. Sõnavara mitmekesisus

Sõnavara mitmekesisus (inglise keeles *Type-Token Ratio* ehk *TTR*) väljendab sõnavara varieerumist tekstis. See saadakse jagades unikaalsete sõnade arvu ehk teksti sõnavara suuruse kõikide tekstis esinevate sõnade arvuga. Sõnavara mitmekesisust on kasutatud tekstide formaalsuse määramiseks, kus madal mitmekesisus näitab formaalsust ja kõrge mitteformaalsust (Sheikha, Inkpen 2012: 16). Hinnangukorpuse alusel näib aga eesti keele puhul olevat olukord vastupidine, kus kõrge mitmekesisus näitab formaalsust ja madal mitteformaalsust, nagu on näha joonisel 1.

Spontaansuse osas olen oma bakalaureusetöös teinud intuitiivse eelduse, kuna ei ole leidnud varasematest uurimustest ei kinnitust ega ka vastuväiteid, mis sai kinnitust juhumetsade abil tunnuste kaale vaadates, sest tunnus osutus spontaansuse hindamisel oluliseks. Lisaks eeldasin, et sõnavara mitmekesisus mõjutab tekstide spontaansust vastupidiselt formaalsusele, kus kõrge mitmekesisus näitab mittespontaansust ja madal spontaansust. See eeldus sai kinnitust vaadeldes hinnangukorpust (joonis 1).

Sõnavara mitmekesisuse arvutan teksti lemmasid vaadates. Kõikide unikaalsete lemmade arvu jagan kõikide lemmade arvuga ja tunnuseks on saadud arv. Seejärel teisendan tunnuse väärtuse punktideks Ühendkorpuse valimi alusel, kus keskmine väärtus on neutraalne ehk 0 punkti ja mida kaugemale jääb see arv keskmisest väärtusest, seda erinevam on punktide arv nullist. Formaalsuse puhul on keskmisest suurem mitmekesisus skoorile positiivse mõjuga ning keskmisest madalam mitmekesisus negatiivse mõjuga. Spontaansus on vastupidine, keskmisest suurem mitmekesisus mõjutab skoori negatiivselt ja madalam positiivselt. Lisadesse olen pannud tabeli (tabel 3), mis näitab, kuidas sõnavara mitmekesisuse väärtus mõjutab dimensioonide skoore.

Tunnuse mõju suunda kontrollisin kasutades hinnangukorpust, kus vaatasin, kuidas on dimensiooni hinnangud seoses tunnuse väärtusega. Kui joonisel olev regressioonijoon tõuseb, on tegu positiivse suunaga, ja kui regressioonijoon langeb, on tegu negatiivse suunaga tunnusega. Joonis 1 kujutab kahte graafikut formaalsuse ja spontaansuse dimensioonide suhetest sõnavara mitmekesisusega. Jooniselt on näha, et formaalsuse hindamisel on sõnavara mitmekesisus positiivse suunaga ja spontaansuse hindamisel negatiivse suunaga.



Joonis 1. **Formaalsuse ja spontaansuse dimensioonide suhted sõnavara mitmekesisusega.**

3.1.2. Keskmine lemmapikkus

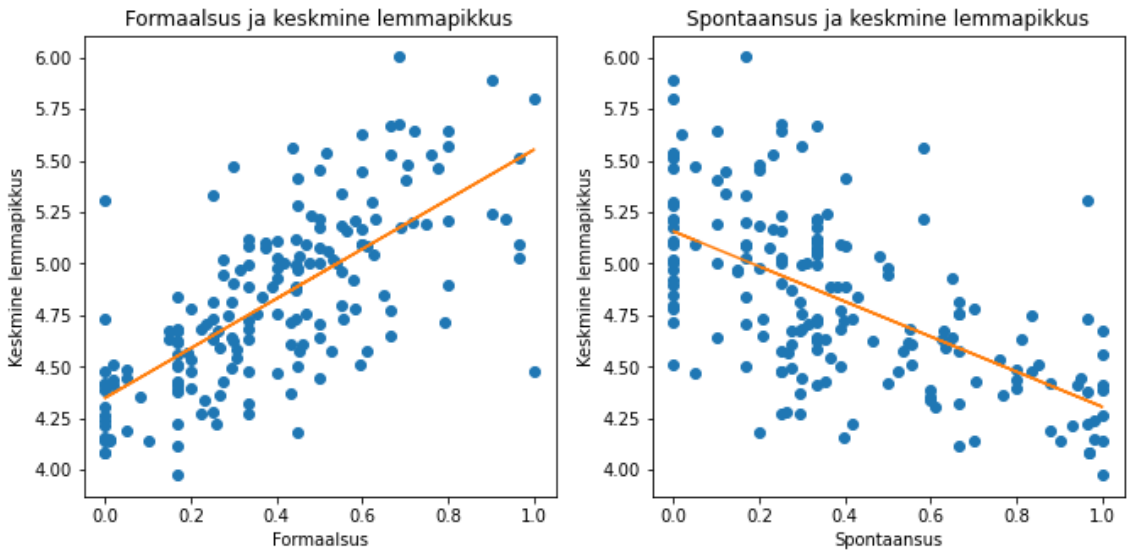
Keskmine sõnapikkus ehk tekstis leiduvate sõnade pikkuste aritmeetiline keskmine on tunnus, mida on kasutatud ingliskeelsete tekstide formaalsuse määramisel ning mis on osutunud selle jaoks määravaks (Sheikha, Inkpen 2012). Spontaansuse puhul tegin bakalaureusetöös siingi intuitiivse eelduse, kuna tunnust ei ole varasemates uurimustes käsitletud. Ka see sai juhumetsade abil kinnitust, sest tunnuse kaal osutus spontaansuse hindamisel väga suureks.

Eesti keeles on liitsõnade moodustamine väga produktiivne ja levinud, seepärast eristan pikki lihtsõnu ja liitsõnu, mis võivad koosneda lühikestest ning lihtsatest osasõnadest. Teen seda käsitledes iga liitsõna osasõna individuaalse lihtsõnana kasutades EstNLTK-

d[6]. Lisaks kasutan tekstis esinevate sõnavormide asemel nende lemmasid ehk algvorme, et käände- ja pöördelõpud ei oleks võetud arvesse. Liidan kokku kõikide tekstis esinevate lemmade tähtede arvud ning jagan selle kõigi tekstis esinevate lemmade arvuga, saades lemmade pikkuste aritmeetilise keskmise.

Siingi kasutan Ühendkorpuse valimi tekstide keskmist, et teisendada tunnuse väärtus skoori punktideks. Mõlema dimensiooni puhul on keskmine lemmapikkus neutraalne, seega väärt 0 punkti. Formaalsuse puhul on keskmisest suurem lemmapikkus skoorile positiivse mõjuga ning madalam negatiivse mõjuga. Spontaansuse puhul on vastupidiselt keskmisest madalam keskmine lemmapikkus positiivse mõjuga ja suurem negatiivse mõjuga. Need mõjud on arvulisel kujul lisades tabelis 4.

Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 2 kujutab formaalsuse ja spontaansuse dimensioonide suhteid keskmise lemmapikkusega. Jooniselt on näha, et formaalsuse hindamisel on keskmine lemmapikkus positiivse suunaga ja spontaansuse hindamisel negatiivse suunaga.



Joonis 2. **Formaalsuse ja spontaansuse dimensioonide suhted keskmise lemmapikkusega.**

[6] EstNLTK. https://estnltk.github.io/

3.1.3. Emotikonid

Emotikonid kui viis kirjutaja emotsioone efektiivselt väljendada on väga levinud uue meedia tekstides, eriti jututubades ja teistes veebi kaudu suhtlemise keskkondades (Muischnek jt 2011: 116). Seepärast on tegu ühe spontaansuse tunnusega. Lisaks on emotikonid ka mitteformaalsuse tunnuseks, kuna need väljendavad emotsioone ja seekaudu subjektiivsust (Sheikha, Inkpen 2012: 7).

Netitekstides leiduvaid emotikone saab jagada neljaks: Unicode'i standardis[7] leiduvad ühe sümboli pikkused pildilised emojid (nagu 😊 ja ☕); foorumile või muule leheküljele omased emotikonid, mis on Ühendkorpusesse salvestatud kui kahe kooloni vahele kirjutatud nimi (näiteks *:kringel:*); kaomojid, ehk mitmest sümbolist koosnevad emotikonid, mis kasutavad kas harva esinevaid või ladina kirjale mitte omaseid sümboleid (nagu ¯_(ツ)_/¯ ja (͡° ͜ʖ ͡°)); lühikesed tähtedest ja/või kirjavahemärkidest koosnevad emotikonid (nagu :) ja T_T). Nimetan kõiki ühiselt emotikonideks. Mitmesümbolilised emotikonid kogusin bakalaureusetööks (Gailit 2021) kahest allikast, Wikipedia[8] ja Looks.wtf[9].

Emotikone kui tunnust vaatlen kolme meetodit kasutades. Esiteks kasutan regulaaravaldisi, et leida kahe kooloni vahelised emotikonid, kuid jätan välja levinud samale formaadile vastavad sümbolid, nagu linkide sees leiduv *:http:*. Seejärel vaatlen emotikone, mis sisaldavad endas vähemalt ühe märgina ladina kirja tähte (näiteks :D). Arvestan emotikoni ümbrusega, et emotikonidena mitte käsitleda olukordi, kus kirjavahemärk on sõnaga koos, nagu loetelueelne koolon näites „*Mehaanik kirjeldab, milline on ideaalne auto:*“, kus *o:* kattub üllatust väljendava emotikoniga. Viimasena käsitlen loetelusid kasutades ülejäänud emotikone, nii mitmemärgilisi kui ka ühesümbolilisi, kuna nende puhul ei ole sõnadega ühendumine probleemiks.

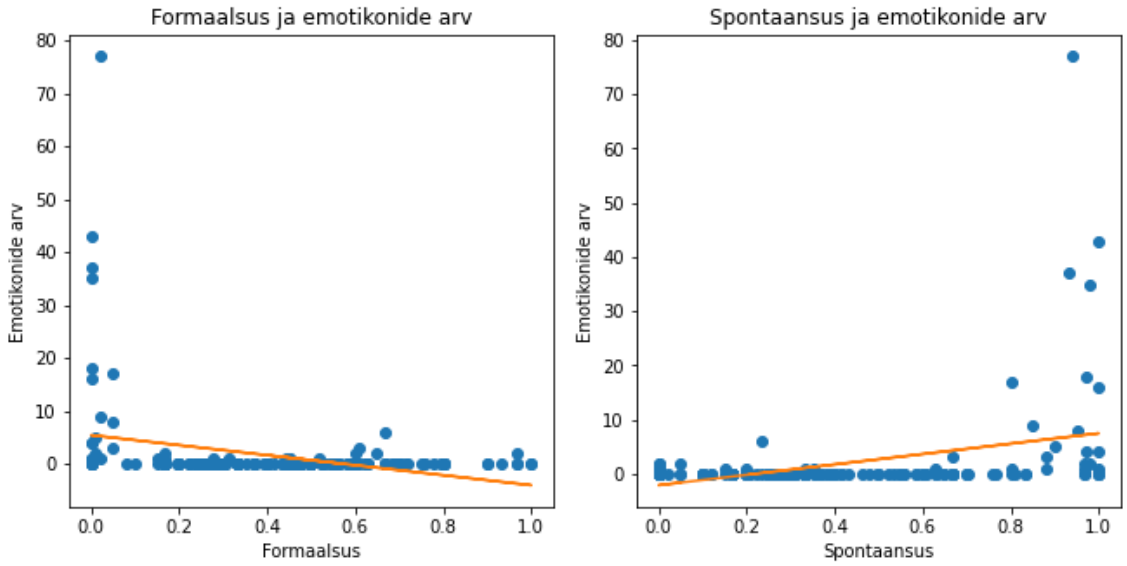
Emotikone vaatlen mudelis kõige esimesena, sest eemaldan need teksti seest teiste tunnuste arvutamise jaoks. Vastasel juhul võib EstNLTK kasutamisel tekkida ootamatuid

[7] **The Unicode Consortium 2021.** Unicode® Emoji Charts v13.1. Mountain View, CA: The Unicode Consortium. http://www.unicode.org/emoji/charts-13.1/
[8] **Wikipedia** https://en.wikipedia.org/wiki/List_of_emoticons. Emotikonid kogutud 05.04.2021.
[9] **Looks.wtf.** https://looks.wtf/. Emotikonid kogutud 06.04.2021.

tulemusi, näiteks emotikonides leiduvad kirjavahemärgid ei pruugi eristuda grammatilisi funktsioone väljendavatest kirjavahemärkidest ning emotikonides sisalduvad tähed eraldatakse kirjavahemärkidest ning käsitletakse nagu kõiki teisi sõnu, mis näiteks vähendaks teksti keskmist lemmapikkust.

Ühendkorpuse valimis leidus emotikone vähem kui pooltes tekstides, seega tunnusena vaatlen ainult, kas emotikone esineb. Kui neid esineb, siis spontaansuse skoor tõuseb ning formaalsuse skoor langeb maksimumväärtuse ehk viie võrra. Kui emotikone ei esine, siis kummagi dimensiooni skoor ei muutu, kuna tegu on kõige sagedasema ehk neutraalse väärtusega.

Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 3 kujutab formaalsuse ja spontaansuse dimensioonide suhteid emotikonide arvuga. Jooniselt on näha, et formaalsuse hindamisel on emotikonide arv negatiivse suunaga ja spontaansuse hindamisel positiivse suunaga.



Joonis 3. **Formaalsuse ja spontaansuse dimensioonide suhted emotikonide arvuga.**

3.1.4. Käändsõnade protsent

Lepingutekstid kui stereotüüpsed formaalsed tekstid sisaldavad rohkelt pikki, rohketest nominalisatsioonidest ehk nimisõnafraasidest koosnevaid lauseid (Reinsalu 2011: 232−233). Seega, mida rohkem esineb tekstis nimisõnu ja nende laiendeid, seda

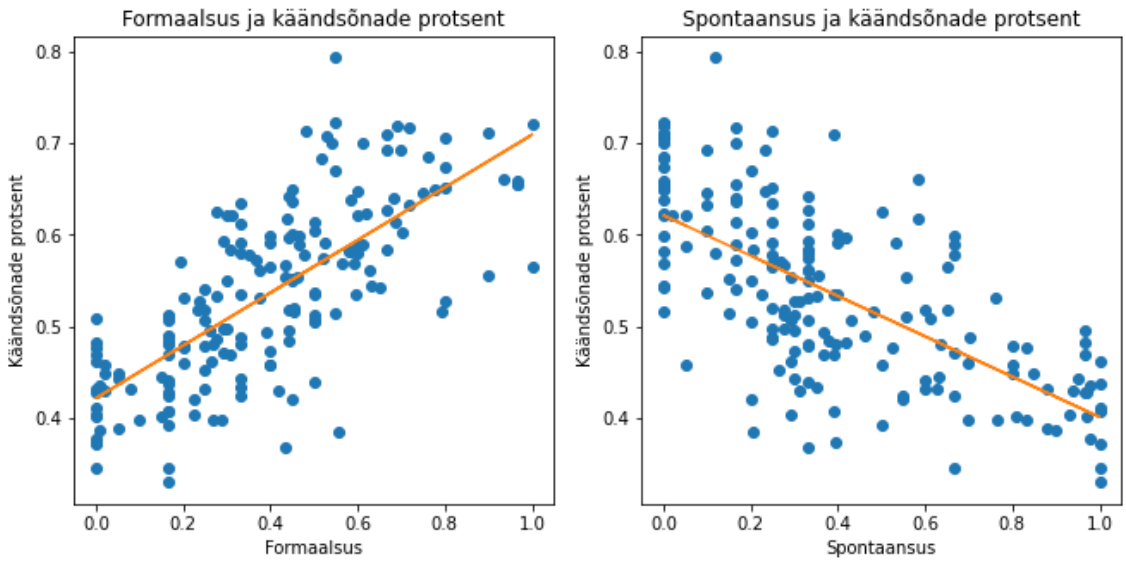
formaalsem on tekst. Seevastu verbid, määrsõnad, asesõnad ja asemäärsõnad vähendavad teksti formaalsust, suurendades selle mitmetimõistetavust ja kontekstuaalsust. (Kerge, Pajupuu 2010: 384–385)

Käändsõnade protsenti kui spontaansuse tunnust ma bakalaureusetöös ei käsitlenud, kuna varasemates uurimustes seda ei oldud täheldatud ning ma ei osanud arvata, et tegu võiks olla spontaansuse jaoks olulise tunnusega. Juhumetsadega spontaansuse tunnuste kaale vaadates katsetasin ka selle tunnuse lisamist, mis osutus väga edukaks, sest käändsõnade protsent osutus vaadeldavate tunnuste seast spontaansuse hindamisel kõige olulisemaks.

Tunnust arvutasin kasutades sõnaliike, kus otsisin nimisõnu, pärisnimesid, omadussõnu, kaassõnu ja arvsõnu. Arvutasin käändsõnade osakaalu kõikidest teksti sõnadest jagades käändsõnade arvu kõikide tekstis leiduvate sõnade arvuga. Tegin seda, et teksti pikkus ei mõjutaks tunnuse väärtust.

Ühendkorpuse valimist leidsin keskmise käändsõnade protsendi, mille määrasin neutraalseks ehk nulliks. Sellest madalamad protsendid said positiivse mõju spontaansuse skoorile ja negatiivse mõju formaalsuse skoorile ning suuremad protsendid vastupidi, negatiivse mõju spontaansuse skoorile ning positiivse mõju formaalsuse skoorile. Need mõjud on arvulisel kujul lisades tabelis 5.

Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 4 kujutab formaalsuse ja spontaansuse dimensioonide suhteid käändsõnade protsendiga. Jooniselt on näha, et formaalsuse hindamisel on käändsõnade protsent positiivse suunaga ja spontaansuse hindamisel negatiivse suunaga.



Joonis 4. **Formaalsuse ja spontaansuse dimensioonide suhted käändsõnade protsendiga.**

3.1.5. Isikulised asesõnad ja verbid

Teksti formaalsuse uurimisel on täheldatud, et nii asesõnade kui ka verbide isikud tähistavad, kui formaalne või mitteformaalne on tekst. Kolmas isik väljendab tekstis formaalsust, nii asesõnade kui ka verbide puhul. Seevastu esimene ja teine isik väljendavad üldiselt mitteformaalsust, kuid oluliseks erandiks on teise isiku puhul kasutatavad formaalsed asesõnad. (Sheikha, Inkpen 2012: 15)

Spontaansuse puhul on aga tunnusena täheldatud ainult esimese isiku asesõnade rohkust. Need esinevad suulistes narratiivides tihti, kuna kõnelejad viitavad endale sagedamini kui kirjalikes narratiivides. (Lindström, Toomet 2000: 178–179) On eeldatav, et see esineb ka väga spontaansetes kirjalikes tekstides, mis peegeldavad rohkelt suulise keele omadusi.

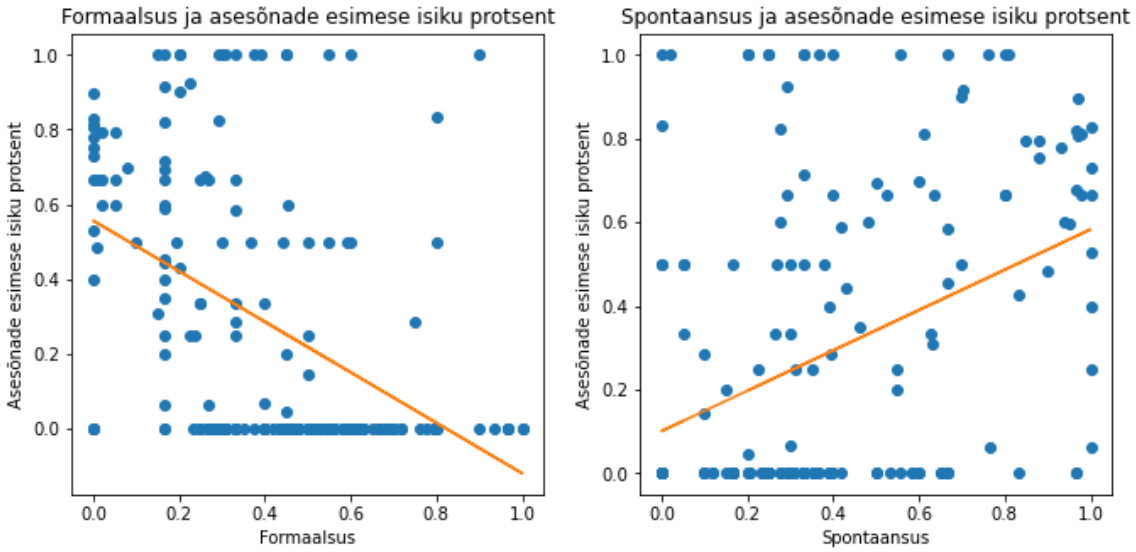
Nii isikulisi asesõnu kui ka verbivorme olen vaadelnud kasutades EstNLTK morfoloogilist analüsaatorit. Isikuliste asesõnade puhul võtan esmalt tekstist välja kõik asesõna märgendi saanud sõnad. Seejärel vaatlen nende lemmasid ning arvutan, kui palju iga isiku kohta vastavaid asesõnu leidub. Tunnuseks arvutan, kui palju iga isiku asesõnu leidub võrreldes kõikide asesõnadega tekstis, jagades isiku asesõnade arvu kõikide asesõnade arvuga tekstis.

Verbide puhul vaatan nende lõpuformatiive, mida EstNLTK tagastab vormi analüüsina. Käsitlen ainult formatiive, mis väljendavad ainult üht kindlat isikut, eemaldades mitut isikut kui ka isikut mitte väljendavad verbivormid. Ka siin arvutan tunnuseks iga isiku verbide protsendi kõikidest verbidest tekstis.

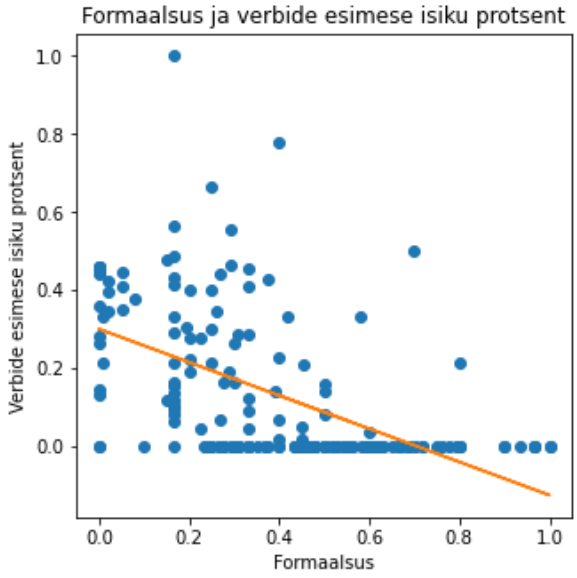
Täheldaksin aga, et ma käsitlen kõiki teise isiku mitmuse esinemisi kui ainult mitmust, ignoreerides teietamist ehk selle kasutust formaalse teise isiku ainsuse asesõnana. Teen seda, kuna verbide puhul ei ole võimalik programmiliselt eristada, kumma kasutusega on tegu, ning asesõnadegi puhul ei pruugi formaalse kasutuse suur algustäht olla eristatav, näiteks lause alguses.

Formaalsuse skoori mõjutavad esimese ja teise isiku asesõnad ja verbid negatiivselt, kus mida rohkem neid on, seda suuremal määral vähendavad need skoori. Kolmanda isiku asesõnad ja verbid mõjutavad skoori positiivselt, suurendades skoori seda rohkem, mida suurem nende tekstis esinemise protsent on. Spontaansust mõjutab ainult esimese isiku asesõnade protsent, kus mida suurem see on, seda kõrgem on teksti spontaansuse skoor. Need mõjud on arvulisel kujul lisades tabelites 6 kuni 11.

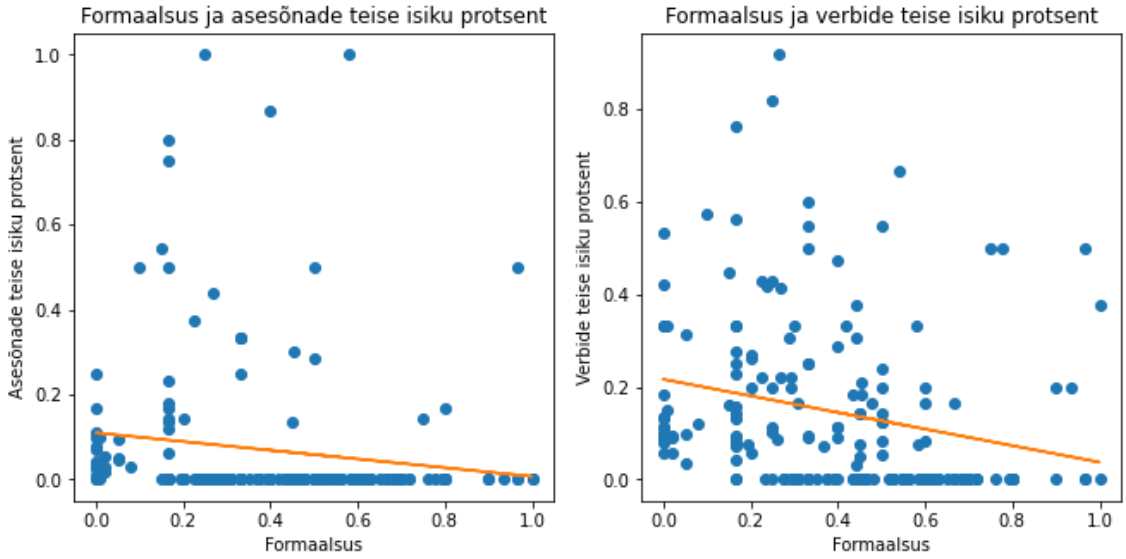
Kontrollisin kõigi kuue tunnuse mõju suunda kasutades hinnangukorpust. Joonis 5 kujutab formaalsuse ja spontaansuse dimensioonide suhteid esimese isiku asesõnade protsendiga. Jooniselt on näha, et formaalsuse hindamisel on esimese isiku asesõnade protsent negatiivse suunaga ja spontaansuse hindamisel positiivse suunaga. Joonis 6 kujutab formaalsuse dimensiooni suhet esimese isiku verbide protsendiga ning joonis 7 kujutab formaalsuse suhet teise isiku verbide ja asesõnade protsentidega. Mõlemad joonised näitavad, et tunnused on negatiivse suunaga. Joonis 8 kujutab formaalsuse suhet kolmanda isiku asesõnade ja verbide protsentidega, mis on mõlemad positiivse suunaga, kuid asesõnade puhul on see väga nõrk. See võib olla tingitud vaadeldava andmestiku väiksusest või viidata, et tunnuse kaal on madal. Olen seosed välja toonud ainult siis, kui olen kasutanud tunnust dimensiooni hindamisel, mistõttu olen spontaansuse puhul välja toonud ainult esimese isiku asesõnade protsendi.



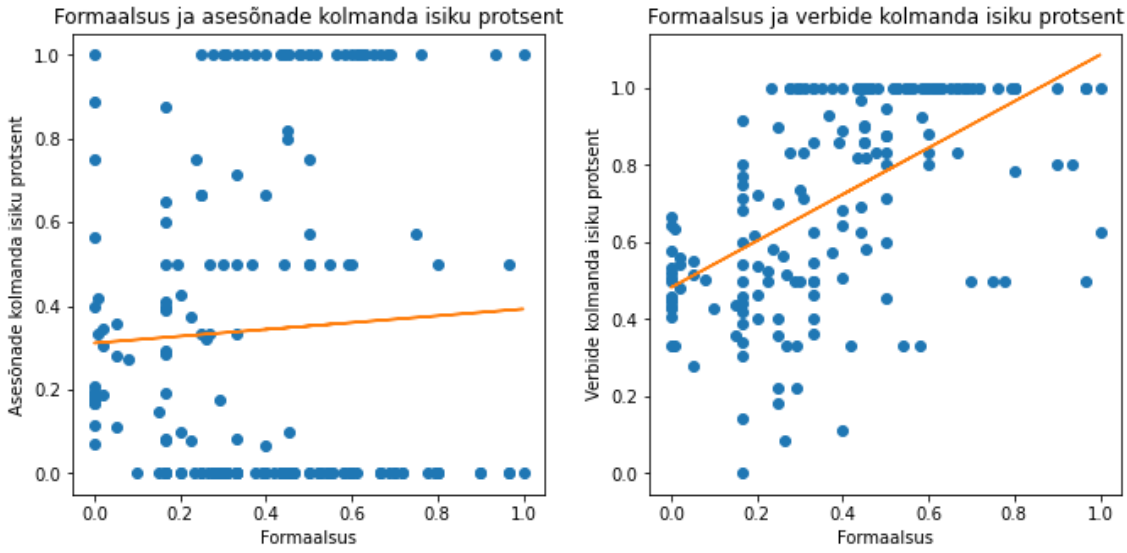
Joonis 5. **Formaalsuse ja spontaansuse dimensioonide suhted esimese isiku asesõnade protsendiga.**



Joonis 6. **Formaalsuse suhe esimese isiku verbide protsendiga.**



Joonis 7. **Formaalsuse dimensioonide suhted teise isiku asesõnade ja verbide protsentidega.**



Joonis 8. **Formaalsuse dimensioonide suhted kolmanda isiku asesõnade ja verbide protsentidega.**

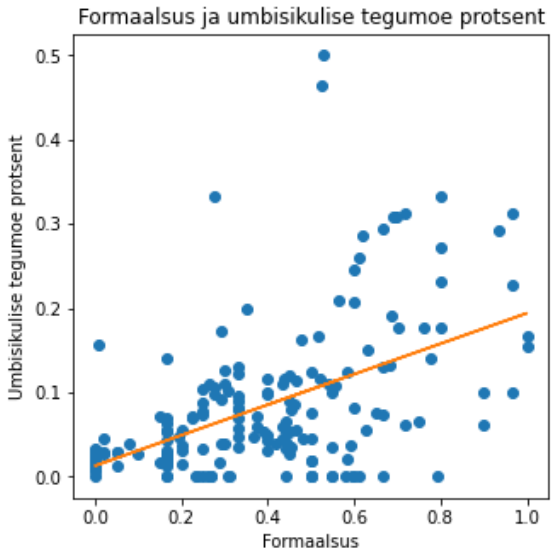
3.1.6. Umbisikuline tegumood

Impersonaal ja teised tegevussubjekti varjamise või tahaplaanile jätmise keelelised meetodid on varasemalt täheldatud ühe formaalsuse tunnusena tekstides. Vastupidiselt on isikuline tegumood ning teised tegevussubjekti väljatoomise meetodid mitteformaalsust väljendavad. (Sheikha, Inkpen 2012: 16)

Eesti keeles on tegevussubjekti varjamiseks kõige tavalisem vahend impersonaali ehk umbisikulise tegumoe kasutamine. See on ka ainus vahend, mida oma mudelis vaatlen. Kasutades EstNLTK morfanalüsaatorit, leian arvu verbidest, mille puhul on vähemalt üks morfoloogiline tõlgendus umbisikuline. Seejärel jagan impersonaalsete verbide arvu kõikide verbide arvuga, saades tunnuse kui protsendi.

Formaalsuse skoori mõjutab umbisikulise tegumoe protsent positiivselt, kus mida kõrgem on protsent, seda tugevam on tunnuse mõju. Need mõjud on arvulisel kujul lisades tabelis 12.

Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 9 kujutab formaalsuse dimensiooni suhet umbisikulise tegumoe protsendiga. Jooniselt on näha, et formaalsuse hindamisel on umbisikulise tegumoe protsent positiivse suunaga.



Joonis 9. **Formaalsuse dimensiooni suhe umbisikulise tegumoe protsendiga.**

3.1.7. *nud*-partitsiibi vormis verbide protsent

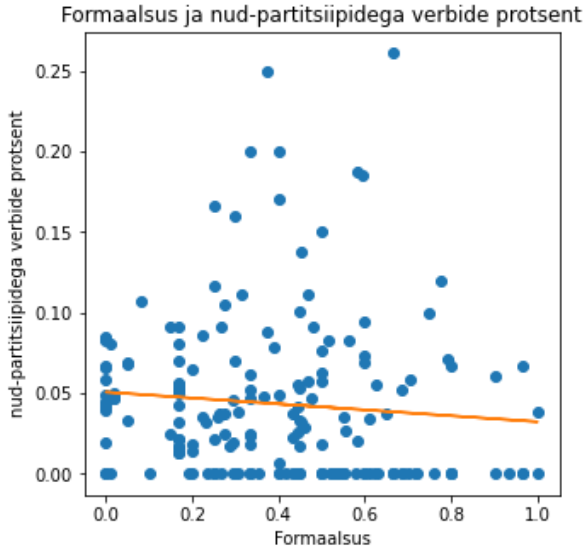
Suulised narratiivid väljendavad vahendatust kasutades suures osas *nud*-partitsiipi (Lindström, Toomet 2000: 181). Kuna spontaansed tekstid on sarnased suulise kõnega, saab öelda, et mida rohkem on tekstis *nud*-partitsiipi, seda vähem formaalsem on tekst.

Arvutan tunnust kasutades EstNLTK morfoloogilist analüüsi, leides arvu kõikidest verbidest, millele on antud *nud*-partitsiibi verbivormi tähis. Seejärel arvutan *nud*-

partitsiibi vormis verbide protsendi, jagades nende arvu kõikide tekstis esinevate verbide arvuga.

Formaalsuse skoori mõjutab tunnus negatiivselt, mistõttu lahutan skoorist seda suurema arvu, mida suurem on *nud*-partitsiipidega verbide protsent tekstis. Need mõjud on arvulisel kujul lisades tabelis 13.

Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 10 kujutab formaalsuse dimensiooni suhet *nud*-partitsiipidega verbide protsendiga. Jooniselt on näha, et formaalsuse hindamisel on *nud*-partitsiipidega verbide protsent negatiivse suunaga, kuigi nõrgalt. See võib olla tingitud vaadeldava andmestiku väiksusest või viidata, et tunnuse kaal on madal.



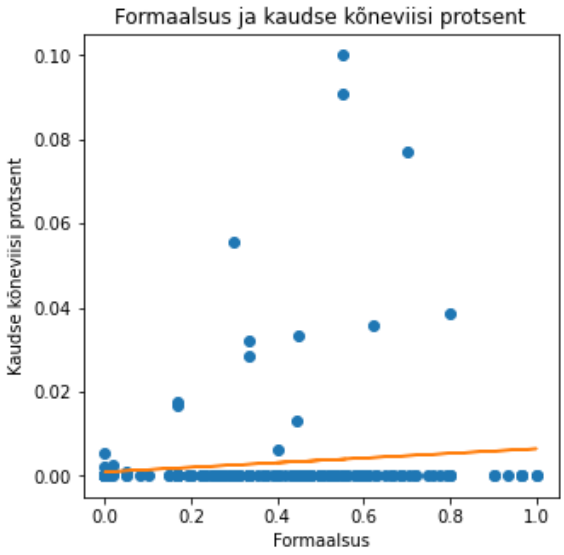
Joonis 10. **Formaalsuse dimensiooni suhe *nud*-partitsiipidega verbide protsendiga.**

3.1.8. Kaudse kõneviisi protsent

Kaudse kõneviisi väljendamise viisidest seostub *vat*-tunnuseline kaudne kõneviis peamiselt formaalsete suhtlussituatsioonidega (Lindström, Toomet 2000). Kõnekeeles kasutatakse lisaks *da*-infinitiivi, kuid seda ei saa mudelis kasutada, kuna kaudse kõneviisi väljendamine ei ole *da*-infinitiivi ainus funktsioon. Seepärast kasutan mudelis vaid *vat*-tunnuselist kaudset kõneviisi.

Arvutan tunnuse väärtuse kasutades EstNLTK morfanalüsaatorit, otsides tegusõnu, millel on kaudse kõneviisi märgend. Seejärel jagan nende arvu kõikide verbide arvuga, saades protsendi. Tunnus mõjutab formaalsust positiivselt ning kuna seda esineb Ühendkorpuse valimis väga vähestes tekstides, alla 16,7% protsendi, vaatan ainult, kas väärtus ületab nulli, seega kas tunnust tekstis esineb. Kui seda esineb, suurendan teksti formaalsuse skoori maksimumväärtuse ehk viie võrra.

Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 11 kujutab formaalsuse dimensiooni suhet kaudse kõneviisi protsendiga. Jooniselt on näha, et formaalsuse hindamisel on kaudse kõneviisi protsent positiivse suunaga, kuigi pigem nõrgalt. See võib olla tingitud vaadeldava andmestiku väiksusest või viidata, et tunnuse kaal on madal.



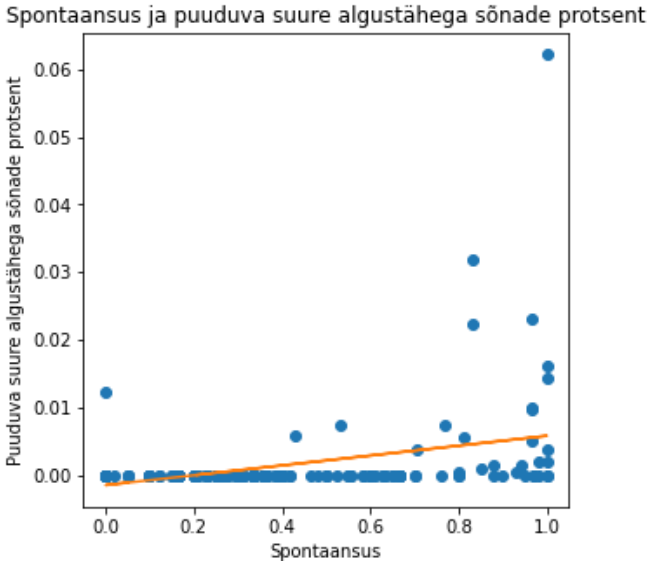
Joonis 11. **Formaalsuse dimensiooni suhe kaudse kõneviisi protsendiga.**

3.1.9. Puuduva suure algustähega sõnade protsent

Internetis on spontaansetes suhtluskeskkondades, sealhulgas sotsiaalmeedias ja jututubades, levinud mitmete normeeritud kirjakeele suur- ja väiketähe reeglite mitte järgimine. Üks selle väljendumistest on lausete ja pärisnimede alustamine suurtähe asemel väiketähega. (Muischnek jt 2011: 117–119)

Otsin tekstidest selliseid sõnu kasutades EstNLTK lausestajat, mis jagab tekstid lauseteks, ning morfoloogilist analüsaatorit, et leida pärisnimede sõnaliigi märgendi saanud sõnad. Tunnuseks arvutan puuduvate suurte algustähtede protsendi, jagades väikese algustähega lausealguliste sõnade ja väikese algustähega pärisnimede arvude summa kõikide lausete ja pärisnimede summaga, saades seega protsendi sõnadest, millel puudub vajalik suur algustäht.

Kuna Ühendkorpuse valimis on puuduvate suurte algustähtedega tekste alla veerandi, vaatan spontaansuse skoori suurendamiseks ainult, et tunnus tekstis esineks. Kui seda esineb, siis spontaansuse skoor tõuseb maksimumväärtuse ehk viie võrra. Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 12 kujutab spontaansuse dimensiooni suhet puuduva suure algustähega sõnade protsendiga. Jooniselt on näha, et spontaansuse hindamisel on puuduva suure algustähega sõnade protsent positiivse suunaga.



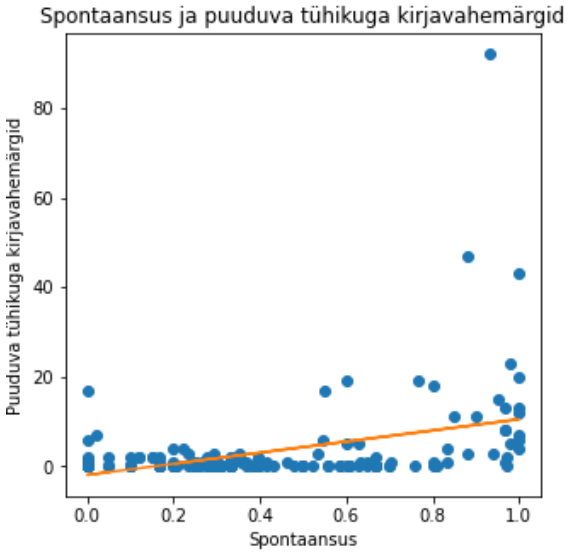
Joonis 12. **Spontaansuse dimensiooni suhe puuduva suure algustähega sõnade protsendiga.**

3.1.10. Puuduva tühikuga kirjavahemärgid

Uue meedia tekstides, nagu blogipostitustes ja jututubades, on väga sage kirjavigade tegemine. Üks levinud kirjavigade tüüp on kirjavahemärkide kirjutamine ilma tühikuteta, näiteks puudub tühik koma järel fraasis „*maja,kus see toimus*“.

Otsin selliseid olukordi kasutades regulaaravaldisi. Vaatlen kirjavahemärke, mis esinevad kahe sõna vahel ilma tühikuta. Erandiks on aga sidekriips, mida tunnuse väärtust arvutades ei käsitleta. Selle põhjuseks on asjaolu, et sidekriips kui kirjavahemärk võib esineda õigekirjareeglite alusel sõnade sees, näiteks liitsõna „*jää-äär*“ ja määrsõna „*tipa-tapa*“.

Ühendkorpuse valimis on puuduvate tühikutega kirjavahemärke alla poolte tekstide, seepärast vaatlen spontaansuse skoori suurendamiseks ainult, et tunnust tekstis esineks. Kui seda esineb, siis spontaansuse skoor tõuseb maksimumväärtuse ehk viie võrra. Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 13 kujutab spontaansuse dimensiooni suhet puuduva tühikuga kirjavahemärkide arvuga. Jooniselt on näha, et spontaansuse hindamisel on puuduva tühikuga kirjavahemärkide arv positiivse suunaga.



Joonis 13. **Spontaansuse dimensiooni suhe puuduva tühikuga kirjavahemärkide arvuga.**

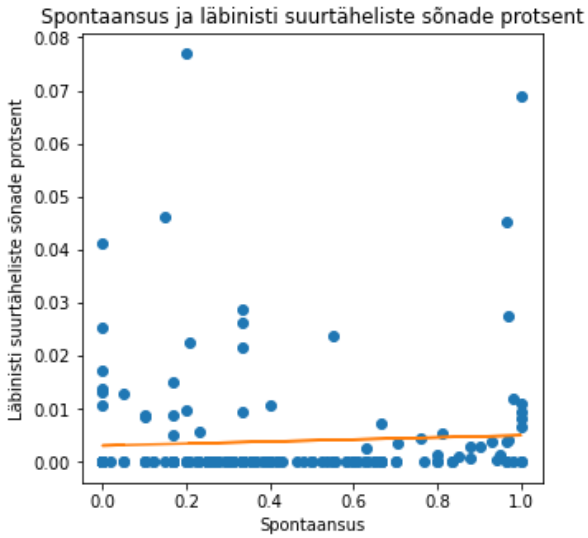
3.1.11. Läbinisti suurtäheliste sõnade protsent

Veel üks viis normeeritud kirjakeele suur- ja väiketähe reegleid spontaansetes kirjalikes suhtluskeskkondades mitte järgida on suurtähtede kasutamine rõhutamise või emotsioonide väljendamise eesmärgil (Muischnek jt 2011: 117–119).

Selliste sõnade leidmiseks vaatlen kõiki vähemalt kahe tähemärgi pikkuseid sõnu, mis oleksid läbivalt suurte tähtedega kirjutatud. Seejärel kontrollin EstNLTK abil, et sõnad ei oleks saanud lühendi märgendit, sest leidub mitmeid lühendeid, mida tuleb kirjutada suurtähtedega (nagu *MTÜ*) ja seega ei saa neid arvestada õigekirjareegleid rikkuvateks. Tunnuse kui protsendi arvutamiseks jagan leitud sõnade arvu kõikide tekstis esinevate sõnade arvuga.

Kui tunnust tekstis ei esine, ei ole teksti spontaansuse skoor mõjutatud, ning kui protsent on nullist suurem, siis mida suurem on protsent, seda suuremal määral on skoor mõjutatud. Need mõjud on arvulisel kujul lisades tabelis 14.

Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 14 kujutab spontaansuse dimensiooni suhet läbinisti suurtäheliste sõnade protsendiga. Jooniselt on näha, et spontaansuse hindamisel on läbinisti suurtäheliste sõnade protsent väga nõrgalt positiivse suunaga. See võib olla tingitud vaadeldava andmestiku väiksusest või viidata, et tunnuse kaal on madal.



Joonis 14. **Spontaansuse dimensiooni suhe läbinisti suurtäheliste sõnade protsendiga.**

3.1.12. Sõnasiseste korduste arv

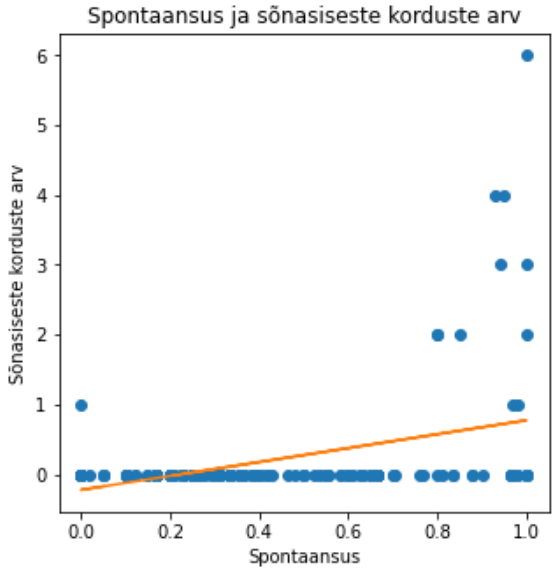
Uue meedia tekstides (nagu blogides, jututubade vestlustes ja sotsiaalmeedia postitustes) on sage rõhutamise eesmärgil korrata sõnade sees üksikuid tähemärke või ka pikemaid

üksuseid (Muischnek jt 2011: 117). See on näiteks tähe *i* kordus sõnes *eiiiiiii* ja *ai* kordus kirjutades *aiaiai*.

Otsin sõnasiseseid korduseid regulaaravaldiste abil, kuid käsitlen ühe tähe ja mitme erineva tähe kordusi erinevalt. Ühe tähe korduseid otsides vaatan, et täht korduks vähemalt neli korda, kuna leidub olukordi, kus kolmekordne kordus on reeglipärane ja seega ebasobiv tunnuse jaoks, näiteks *kukkki*. Neljakordseid tähekordusi õigekirja norme järgivates tekstides seevastu reeglina ei leidu, sest liitsõnad nõuavad kolme ja nelja tähe kordumisel sidekriipsu sõnapiirile lisamist, näiteks *jää-äär*. Mitme tähemärgi kordused peavad olema vähemalt kolm korda korduvad, kuna kahekordsed kordused võivad esineda õigekirja reegleid järgivates tekstides. Käsitlen mõlemat tüüpi sõnasiseseid korduseid koos ühe tunnusena.

Ühendkorpuse valimis esineb sõnasiseseid kordusi alla poolte tekstide, mille tõttu vaatan ainult, kas tunnust tekstis esineb, et suurendada spontaansuse skoori. Kui seda esineb, suurendan teksti spontaansuse skoori maksimumväärtuse ehk viie võrra.

Kontrollisin tunnuse mõju suunda kasutades hinnangukorpust. Joonis 15 kujutab spontaansuse dimensiooni suhet sõnasiseste korduste arvuga. Jooniselt on näha, et spontaansuse hindamisel on läbinisti sõnasiseste korduste arv positiivse suunaga.



Joonis 15. **Spontaansuse dimensiooni suhe sõnasiseste korduste arvuga.**

3.1.13. Eemaldatud tunnused

Bakalaureusetöö (Gailit 2021) raames loodud mudel kasutas eelnevalt loetud tunnustele lisaks mõnda tunnust, mida olen otsustanud magistritöö mudelis mitte kasutada. Kaks tunnust said eemaldatud leksikonide lisamise tõttu. Esimene on EstNLTK oletamiseta morfoloogilisel analüüsil tundmatu sõna analüüsi saanud lühikeste sõnade protsent kui spontaansuse tunnus, kuna tundmatuks jäänud sõnu sai kasutatud leksikonide koostamisel ja seega oleks tegu kattuva tunnusega. Teine eemaldatud tunnus on kirjavigadega sõnade protsent, kuna olen lisanud paljud sagedased kirjavigadega sõnavormid kasutatud leksikonidesse.

Lisaks eemaldasin tunnuste hulgast tajuverbide protsendi kõikidest verbidest ja korduvate sõnade arvu. Mõlema tunnuse puhul osutus nende kaalude arvutamisel, et need mõjutasid juhumetsade keskmise veamäära tõusu negatiivselt ehk kui tunnus juhumetsast eemaldada, keskmine veamäär langes. See näitas, et need tunnused ei olnud mitte ainult ebaolulised, vaid lausa tegid juhumetsa kehvemaks, tõstes keskmist veamäära.

3.2. Leksikonid

On teada, et sõnavara on oluline tunnus teksti liigi määramisel. Seda kinnitab asjaolu, et eelnevalt mainitud Rezapour Asheghi jt (2014) mudel, mis liigitab tekste kasutades ainult tähemärkide *n*-gramme, saab täpsuseks 78,88%, mis on selle ülesande puhul hea tulemus. Selle alusel võib eeldada, et sõnavara vaatlemine oleks kasulik ka tekstide formaalsuse ja spontaansuse dimensioonide hindamisel, et suur kogus tekstides leiduvat informatsiooni ei jääks kasutamata.

Erinevalt aga eelnevalt mainitud mudelitest, ei ole minu töös hetkel võimalik kasutada masinõpet. See eeldab heade tulemuste saavutamiseks suurt korpust, kuid dimensioonide suhtes annoteeritud hinnangute korpus on väike. Seepärast kasutan sõnavara mõju hindamiseks hoopis leksikone, mis sisaldavad tekstide formaalsust ja spontaansust mõjutavaid sõnu ja sõnavorme.

Koostasin leksikonid käsitsi, kasutades EstNLTK oletamiseta morfoloogilist analüsaatorit. Olin bakalaureusetöös (Gailit 2021: 16) eeldanud, et lühikesed tundmatu

sõna analüüsi saanud sõnad ehk sõnad, mis ei esine EstNLTK sisemises sõnastikus, väljendavad spontaansust. See sisemine sõnastik koosneb Õigekeelsussõnaraamatus esinevatest sõnadest ja nendest moodustatavatest liitsõnadest ja tuletistest, seega selles mitte esinevad sõnad on kas uued, väga kõnekeelsed, laensõnad või väga spetsiifilised erialaterminid, mis sõnaraamatusse pole jõudnud.

Leksikonid on moodustatud Ühendkorpuse valimi tekstides esinevatest sõnedest ehk tähemärkide, numbrite ja muude sümbolite kombinatsioonidest. Peamiselt on siiski tegu aga sõnade algvormide ja teiste sõnavormidega. Esmalt tegin sagedusloendi kõikidest valimis esinevatest sõnedest, mis on saanud oletamiseta morfoloogilisel analüüsil tundmatu sõna analüüsi. Koostatud loend koosneb 267 839 erinevast tundmatuks jäänud sõnest, kus kõige sagedasem on "•", mida esineb 13 406 korda. Esimene sõne, mis ei ole märk, on sagedusloendis kaheksandal kohal esinev *it*, mida leidub valimis 2200 korda. Sagedusloendi 267 839 sõnest esineb 202 293 sõna ainult ühe korra, millest võib eeldada, et ka valimist väljaspool esineb neid sõnavorme harva ja seega nende mudelisse lisamine annaks vähe lisainformatsiooni. Seepärast vaatan sagedusloendist ainult selliseid sõnesid, mida esineb valimis vähemalt 10 korda. Neid on kokku 8088.

Leksikonidesse valisin sõnesid tähenduse ja konteksti alusel. Konteksti vaatan esimesena: kui sõne esineb ainult spontaansetes ja mitteformaalsetes kontekstides, nagu foorumites ja blogides, siis on see potentsiaalselt spontaanne ja mitteformaalne. Seejärel vaatan sõna tähendust: kuna valimis on rohkelt erialafoorumeid, on paljud tundmatuks jäänud sõnad osa erialakeelest, kas terminoloogia või erialasläng.

Formaalsust ja/või mittespontaansust väljendavaid sõnavorme ma leksikone koostades ei käsitlenud nende vähesuse ja umbmäärasuse tõttu. Mitmed sõnad, mis võisid potentsiaalselt formaalsust väljendada, väljendasid pigem keerukuse dimensiooni. Mitmel juhul osutus pika ja keerulise sõnavormi allikaks muidu kõnekeelt täis vestlus erialafoorumis, mis väljendab, et sõna on pigem neutraalne formaalsuse poolest. Lisaks ei saa kasutada ka viisakusvormeleid ja -sõnu kui formaalsust väljendavat sõnavara, kuna neid leidub nii ilukirjanduses aga ka sagedaselt kõnekeeles, sageli lausa sarkasmina.

Seepärast sisaldavad leksikonid ainult neid sõnavorme, mis väljendavad ainult spontaansust ja/või mitteformaalsust.

3.2.1. Leksikonide kirjeldused

EstNLTK oletamiseta morfoloogilisel analüüsil tundmatu sõna märgendi saanud sõnede loendi alusel sai loodud kaksteist leksikoni, mis kõik väljendavad erinevat tüüpi tekstis esinevaid sõnavorme. Kategooriatesse jaotatud leksikonid koos nende esinemiste arvuga valimis on kättesaadavad mudeli GitHubist[10]. Järgnevalt kirjeldan kõiki leksikone, millistest sõnedest need koosnevad ning mille alusel on need ühte leksikoni lisatud.

3.2.1.1. Toorlaenud

Toorlaenud on sõnad, mis on eesti keelde laenatud mõnest teisest keelest ilma kirjapilti mugandamata. Kuigi *toorlaen* on vastuoluline mõiste, kuna seda sageli kasutatakse negatiivse hinnangu väljendamiseks, kasutan ma seda, et eristada kirjapildis mugandamata laene mugandatud laenudest, mida kirjeldan alampeatükis *laensõnad*.

Toorlaenude leksikoni kuuluvad mitte ainult algvormid, vaid ka käändelised ja pöördelised sõnavormid. Nende puhul on konteksti vaatamine väga oluline, kuna sageli võib tundmatuks jäänud sõna olla erialane termin, mis võib esineda nii mittespontaansetes ja formaalsetes kui ka spontaansetes ja mitteformaalsetes kontekstides. Valimis esineb rohkelt tundmatu sõna analüüsi saanud toorlaene, kuid neid kõiki ei saa erialalise tähenduse tõttu dimensioonide hindamisel kasutada.

Toorlaenud pärinevad valdavalt inglise keelest. Enim leidub vulgaarseid ja sõimusõnu, nagu *fuck*, *shit*, *fucking*, *bitch* ja *bullshit*. Sõna *sex* kuulub samuti siia loendisse, kuid seda saab käsitleda eesti keeles leiduva sõna *seks* märgiasendusena, kus *ks* on asendatud *x*-ga. Leidub palju omadussõnu, sealhulgas *cool*, *extra*, *chill*, *creepy* ja *aussi* (alternatiivne kirjaviis sõnast *Aussie*, mis tähendab *austraalia* või *austraallane*).

Toorlaenude leksikonis esineb palju sotsiaalmeediaga seotud sõnu, nagu *selfie*, *outfiti* ja *youtubes* (*YouTube* seesütlevas käändes). Lisaks on leksikonis mitmeid toorlaene, mis

[10] Link leksikonide kaustale: https://github.com/kgailit/Gailit-2023/tree/main/Loendid/Leksikonid_infoga

pärinevad erialakeelest, näiteks arvutimängude kohta käiv *gameplay* ja reisiblogides *backpacker*.

Leidub ka näiteks prantsuskeelne toorlaen *dejavu*, mis on loendis aga pigem ebakorrektse kirjapildi tõttu. Korrektne vorm *déjà-vu* ning alternatiivsed kirjapildid *déjà vu* ja *deja vu* leksikoni ei kuulu, kuna erinevalt vormist *dejavu* esinesid need nii (mitte)spontaansetes kui ka (mitte)formaalsetes tekstides. Venekeelseid toorlaene leidub valimis vähe, sest eestikeelsetes tekstides on kirillitsa asendatud kõla järgi ladina tähestikuga ja seega käsitlen neid kui harilikke laensõnu.

3.2.1.2. Lühendid

Valimis esineb palju lühendeid, mida ei saa tunnuseks pidada, sealhulgas *EL*, *EU*, *EKRE* ja *spl*. Seega ei kuulu leksikoni ametlikud suurtähtlühendid ning mõõtühikute lühendid, need esinevad nii spontaansetes ja mittespontaansetes kui ka formaalsetes ja mitteformaalsetes tekstides.

Tunnusena käsitletavad lühendid jagunevad kolme kategooriasse:

1. Toorlaenulised lühendid, nagu *wtf*, *aka*, *lol*, *yolo* ja *tnx*;
2. Eestikeelsed sõnaühendite valiktähtlühendid, kus iga täht väljendab erinevat sõna, nagu *jnejne* ja *nkn*;
3. Eestikeelsed üksiksõna valiktähtlühendid, kus on sõnadest välja jäetud üksikud tähed, enamasti vokaalid, nagu *krt*, *vbl* ja *vbla*, *pmst* ja *põmst*, *tglt* ja *ntx*. Seda kategooriat esineb kõige rohkem.

3.2.1.3. Partiklid

Partiklid on lühikesed sõnad, sageli lühenenud sõnavormid. Partikleid on erinevaid tüüpe, mille kõigi tähendused on erinevad. Antud leksikonis sisaldub erinevat tüüpi partikleid:

- Dialoogipartiklid, mis väljendavad kas kuuldel olekut või sõnumi vastuvõtmist;
- Afektiivsed partiklid, mis avaldavad arvamust ja tundeid;
- Aktiivsed suhtluspartiklid, mis püüavad tähelepanu;
- Piiripartiklid, mis tähistavad lausungi algust või lõppu;
- Toimetamispartiklid, mis tähistavad mõttepause. (Hennoste 2002: 71–72)

Käsitlen partiklite all ka laensõnadest, sealhulgas toorlaenudest, partikleid, nagu prantsuse keelest laenatud suhtluspartikkel *voila* (ja ka edasimugandus *voilaa*), inglise piiripartiklid *well* ja *enivei* ning afektiivne *kammoon* ja vene päritolu suhtluspartikkel *tavai* ning afektiivne partikkel *pohh*. Partiklina esinevad lühendid, nagu afektiivne partikkel *irw*, ja suhtluspartiklid *kle* ja *btw*, olen aga paigutanud lühendite leksikoni.

Eestikeelseid partikleid leidub leksikonis palju rohkem. Piiripartiklid on näiteks *vä*, *nu* ja *njah*; afektiivsed partiklid on *hehe* ja *oeh* ning dialoogipartiklid on *aaa*, *mmm*, *ooo*, *hmm*, toimetamispartiklid on näiteks *ää* ja *eee* ning aktiivsed suhtluspartiklid on *jou* ja *oih*. Paljud partiklid esinevad mitmel kujul, peamiselt märgikorduste tõttu: *ää* kõrval esineb ka *äää* ning *haha* võib olla ka *hahaha*. Pikemaid kordusi esines valimis alla kümne korra ja seega neid leksikoni ei lisanud. Nende üles leidmine ja mudelis tunnusena kasutamine on aga juba realiseeritud osana bakalaureusetöö mudelist (Gailit 2021: 18).

3.2.1.4. Laensõnad

Leksikonide tarbeks liigitan laensõnadeks sellised sõnad, mis pärinevad mõnest võõrkeelest, kuid erinevalt toorlaenudest on neid eesti kirjaviisi ja hääldustavade jaoks mugandatud. Kuna leksikonidesse kuuluvad ainult spontaansust ja mitteformaalsust väljendavad sõned, siis jätan välja laensõnalised erialaterminid ja ka vanemad laensõnad, mis ei väljenda spontaansust ja mitteformaalsust. Seega sõnad nagu *infrastruktuur* ja *radiaator* sellesse leksikoni ei kuulu.

Leksikoni kuuluvad aga uuemad laensõnad, mis peamiselt pärinevad inglise või vene keelest. Sellised on näiteks mitmed Internetiga seotud sõnad nagu *googeldama* ja *juutuub* ning nende variandid. Leksikonis on lisaks sõnad nagu *tšill* ja *tsillima*, *tsikk, ruulima*, *megalt* ja *üüber*, mis on kõik tüüpiliselt seotud teismeliste ja noorte keelekasutusega. Leidub ka terminilaadseid laensõnu, nagu *laivis* või *kardiot*, mis aga esinevad ainult blogides, foorumites ja teistes spontaansetes ja mitteformaalsetes tekstides.

3.2.1.5. Omasõnad

Omasõnadeks loen eestikeelseid sõnu, mis ei ole laensõnad või mis on väga vanad laensõnad ja seega tunduvad omased. Lisaks arvestan kõiki sõnu, mis on loodud

kasutades eesti keelele omaseid tuletamismeetodeid, sealhulgas ka uuematest laensõnadest moodustatud sõnu.

Mittelaensõnalised omasõnad on näiteks *kuda*, *aind*, *nigu*, *natu*, *ninnu-nännu*, *kööga*, *nimetet*, *kriban*, *emps* ja *jummala*. Nende hulka kuulub veel lisaks palju tuletatud sõnu, nagu *kõrvakad*, *miskine* ja *tutikas*, palju lühendvorme, nagu *jätsi*, *peiks* ja *väss*, ning mitmeid normeeritud kirjakeeles mittelubatud vorme, nagu *lemmikum* kui sõna *lemmik* võrre.

Laensõnad esinevad samadel viisidel, nagu mittelaensõnad. Väga palju leidub lühendvorme, nagu *digikaga* tähenduses *digikaameraga*, *kiltsa* tähenduses *kilomeetrit*, *bensu* tähenduses *bensiini*, *inffi* tähenduses *informatsiooni*, *absull* tähenduses *absoluutselt*, *akva* tähenduses *akvaarium*, *minti* tähenduses *minutit*, *motti* tähenduses *motivatsiooni* ning *prossa* tähenduses *protsenti*. Leidub ka teistsuguseid tuletusliiteid, nagu *kas*-tuletusliide sõnades *radikas* ja *turistikas*.

3.2.1.6. Kokkukirjutamised

Kokkukirjutamised on sõned, kus on kirjakeele normidele vastuoluliselt kokku liidetud kaks sõna üheks sõneks. Mõned kokkukirjutamised on väga levinud, nagu *igalpool*, *misiganes* ja *igastahes*, mille tähendusel võib vahel olla veidi erinev nüanss võrreldes lahku kirjutatud fraasiga. Paljud kokkukirjutamised on aga vaid näpuvead, nagu *ärakasutada*, *mittekunagi*, *minuarvates* ja *ausaltöeldes*.

3.2.1.7. Märgiasendused

Märgiasendused on väga tüüpilised foorumite ja jututubade vestlustes (Muischnek jt 2011: 117). Leksikonis esineb kahte tüüpi märgiasendusi: keelemängu ja täpitäheasendusi. Keelemäng on olukord, kus üks või mitu tähte asendatakse teise märgiga, mis on kas visuaalselt või kõlaliselt sarnane. Sellised on näiteks number *2* asendamaks tähte *ä*, *y* asendamaks tähte *ü*, *x* asendamaks ühendit *ks* ja *6* asendamaks tähte *õ* või vahel ka tähte *ö*, mis on enamasti asendatud *8*-ga.

Täpitäheasendused on teine märgiasenduse viis. Erinevalt keelemängust, mis vahel nõuab teadmisi, mis millega asendub, põhineb täpitäheasendus ainult täpitähtedelt täppide ära

võtmisega. Seega *ü* on *u*, *ö* on *o* ja *ä* on *a*. Selle termini all käsitlen ma ka teisi sarnaseid tähti, nagu *õ* asendajaga *o*. Seevastu *š* asendajaga *s* ja *ž* asendajaga *z* on pigem kirjavead, mis tuleneb välja nende kasutuste kontekstidest.

3.2.1.8. Märgikordused

Märgikordustega sõnad on sõnad, kus ühte tähte esineb rõhu eesmärgil rohkem kui ette nähtud. Väga sage on see sõna *nii* puhul, kus leksikonis on vormid *niii*, *niiii*, *niiiii*, *niiiiii* ja edasi kuni kümne *i*-ni, mille järel esinemiste arv langeb alla piiri. Märgikordused, mis esinevad üle kolme korra, on juba bakalaureusetöö (Gailit 2021: 18) raames loodud mudelis arvesse võetud, kuid kahe tähe pikkuseid kordusi ei arvestatud, kuna eesti keeles on sage, et kaks samasugust häälikut on teineteise kõrval. Seega varasemas mudelis ei ole arvestatud sõnedega nagu *kaa*, *olii* ja *pääris* kui märgikordustega.

3.2.1.9. Keele- ja kirjavead

Käsitlen termini *viga* all kirjakeele normingutele mittevastavaid sõnavorme. Tegu on vastuolulise mõistega, kuna see sageli väljendab negatiivset hinnangut mittestandardse keelekasutuse vastu. Kasutan seda terminit selle lühiduse tõttu.

Keele- ja kirjavead kuuluvad ühte leksikoni, kuna mõlemad vead on sarnased ja sageli on nende vahel keeruline vahet teha. Kui *kellegil*, *millegil* ja teised sarnased vormid on piisavalt sagedased keelevead, et toimub arutelu nende leksikaliseerumise üle, on käände- ja pöördevormides nagu *taodelda*, *hakkata* ja *suhkurt* keerulisem selgeks teha, kas tegu oli kogemata valesti kirjutamisega või mitte.

Siia leksikoni kuuluvad veel teisedki sagedased vead, nagu *mõtetu* ja *mõtekas*, *potensiaali*, *robotide*, *taigent* ja *videode*. Kogemata tekkinud kirjavead, tuntud ka kui näpuvead, on näiteks *peaagu*, *sisi*, *niimodi* ja *nind*. Leksikoni olen lisanud veel *nud*-pöördevormi variandi *-nd*, nagu *väsind* ja *saand*, kuigi seda võib pidada teadlikuks asenduseks. Ka *ž* asendus *z*-ga või *zh*-ga ja *š* asendus *s*-ga või *sh*-ga kuulub siia kategooriasse, nagu sõnad *maneez*, *massaz*, *garaazhis*, *dushi*, *shokolaad* ja *sokolaadi*.

3.2.1.10. Numbrivead

Numbrivead on olukorrad, kus esineb kirjaviga numbriga sõnes. Peamiselt on tegu olukordadega, kus ühik on arvuga kokku liidetud, nagu *30min*, *1tl*, *40mcg* ja *1m2* (tähenduses *1* m^2, kus ühiku ruudu tähis on vormistuse kaotanud korpuse moodustamise protsessis).

3.2.1.11. Formaadivead

Formaadivigadeks pean sõnesid, kus esinevad kindlad märgiasendused, mis on tekkinud teksti vale formaadiga töötlemisel. Formaat tähendab antud kontekstis kodeeringut, millesse tekst on loomisel salvestatud. Kodeeringuid on mitmeid erinevaid, ASCII-st UTF-8-ni. Kui tekst lugeda vales kodeeringus sisse, võivad mitmed märgid asenduda teiste märkidega ning tekivad formaadivead. Korpuste puhul on keeruline kindlaks teha, millisel hetkel formaadiviga tekkis. Palju formaadivigu on teksti allikates, kus lehekülg on vale kodeeringuga Internetti üles laetud, kuid ei saa välistada, et viga tekkis teksti korpusesse lisamisel. Vea tekke põhjuseks võis olla erinevus teksti salvestamisel kasutatud kodeeringu ja teksti algse kodeeringu vahel.

Kuigi formaadivigu ei saa tunnusena kasutada, koostasin nendest siiski leksikoni. Formaadivead mõjutavad peamiselt eesti tähestiku tähti *ä*, *ö*, *ü* ja *õ*. Mõjutatud tekstides võivad need olla asendatud mõne näiliselt suvalise märgijadaga, nagu *vãµi* sõna *või* asemel, või mõne teise sarnase vokaaliga, mis ei leidu eesti tähestikus, nagu *vői* ja *vòi*. Viimase kategooriaga esineb aga lisaprobleem: kuigi sageli on tegu formaadiveaga, siis selliseid vigu võib esineda ka kirjavigadena, nimelt kui teksti kirjutatakse kasutades nutitelefonide klaviatuuri ja valides vale täpitähe. Leksikonis välja toodud vead esinesid tekstides aga süsteemselt ehk erinevalt näpuveast, on kõik tähe esinemised asendatud ühe ja sama veaga.

3.2.1.12. Murdesõnad

Murdesõnad on kirjakeeles enamasti kasutusel mitteformaalsetes ja spontaansetes tekstides stilistilisel eesmärgil. Sellised sõnad on näiteks *hää*, *pääl*, *köige*, *läeb*, *õhtal*, *perra*, *peris* ja *inime*. Kuna korpustes võib leiduda nii murdelisi kui ka kirjakeelseid

tekste, ei saa pidada murdesõnu tunnusteks, kuid katsetamise huvides on nendest moodustatud leksikon.

3.2.2. Probleemid leksikonide koostamisel

Leksikone koostades selgus, et leidub palju sõnesid, mille puhul on keeruline selgeks teha, kas need väljendavad spontaansust ja/või mitteformaalsust. Edasi loetlen esile tõusnud probleeme kategooriate kaupa.

3.2.2.1. Nimede ja sõnade eristamine

Leksikonides leidub sõnesid, mis sõltuvad tähtede suurusest. Suure algustähega, või läbivalt suurte tähtedega mitmete lühendite puhul, väljendavad need nime. Väikese algustähega väljendab sama sõna aga spontaansust ja/või mitteformaalsust. Näiteks partiklit *assa* saab kasutada fraasis *Assa mait!*, või lugeda tootjanimena lukkudelt ja võtmetelt.

See probleem oleks osaliselt lahendatav nii, et igale sõnele leksikonis lisada märge, kas on vaja kontrollida, et see oleks vaid väikeste tähtedega kirjutatud. See aga ei ole ideaalne lahendus, kuna lause alguses on võimatu eristada nime ja spontaansust ja/või mitteformaalsust väljendavat sõne. Samuti jääb informatsioon puudu juhtudel, kui rõhu eesmärgil on spontaansetes tekstides kasutatud läbivalt suuri tähti, näiteks *ASSA MAIT KUI HEA!!*, või kui on stiili eesmärgil kasutatud läbivalt väikeseid tähti. Seepärast otsustasin sellised sõned leksikonidest välja jätta.

3.2.2.2. Paralleelsed eesti mugandused ja originaalsed kirjavormid

Leidub mitmeid laensõnu, kus on samaaegselt kasutusel nii eesti mugandatud variant kui ka selle originaalne, toorlaenuline variant. Sellised paarid on näiteks *pizza* ja *pitsa*, *puzzle* ja *pusle* ning *chia* ja *tšiia* või ka *õlisalvei*. Ühes tekstis võivad esineda mõlemad variandid samaaegselt, isegi kõrvuti. Näiteks mitmes veebitoidupoes on külmutatud toodete loendis nii *pitsad* kui ka *pizzad* ning veebiraamatupoodides on lauamängukategooria *puzzle* aga tootel nimi *pusle*.

3.2.2.3. EKI soovitused

Leidub palju sõnu, mille kohta on Eesti Keele Instituut soovitanud kasutada alternatiivset, eesti keelele omasemat varianti. Neid leidub nii keelenõuvakas kui ka Ametniku soovitussõnastikus. Kuigi tegu on keeleliselt mittesoovitatud variantidega, on need siiski aktiivses kasutuses. Sageli esinevad need toimetamata või vähe toimetatud tekstides, olles näiliselt spontaansuse tunnuseks, kuid vaadates nii mittesoositud kui ka soovitatud varianti, on nende kasutuskontekst sama ja seega ei saa öelda, et üks väljendab tunnust rohkem kui teine.

Sellised sõnad on näiteks *friteerima*, soovituslikult *frittima* ja mille puhul esinevad mõlemad nii retseptides kui ka blogides, ning *optimiseeritud*, soovituslikult *optimeeritud* ja mille puhul esinevad mõlemad infotehnoloogia teemalistes tekstides, foorumitest tootekirjelduste ning juhenditeni.

3.2.2.4. Mitmetähenduslikud sõnad

Leidub mitmeid sõnesid, millel on mitu tähendust, kus mõni väljendab spontaansust ja/või mitteformaalsust ja mõni mitte. Sellised on näiteks *ex* ja *jaad*. *Ex* võib olla märgiasendus sõnast *eks*, toorlaen tähendusega *endine* või osa ladinakeelsetest fraasidest, nagu *ex libris*, *ex machina* ja *ex*. *Jaad* esineb nii rohelise vääriskivi tähenduses, sõna *jääd* märgiasendusena ja partikli *jaa* osastava käändena.

Kuna ainult lemmasid vaadates ei ole võimalik otsustada, millise tähendusega on tegu, olen otsustanud kõik sellised mitmetähenduslikud sõned leksikonidest välja jätta.

3.2.3. Ühendsõnastik leksikoniks

Vaadeldes ainult dimensioonide suhtes annoteeritud hinnangutega korpust selgus, et mainitud kaheteistkümnesse leksikonisse määratud sõnu leidus nimetatud korpuses väga vähe. Järeldasin, et on vaja lisasõnavara, kuid kuna käsitsi loendi koostamine on ajamahukas ning EstNLTK oletamiseta morfoloogiline analüüs käsitleb iga sõnavormi kui eraldi sõna, otsustasin sõnavara juurde lisada EKI Ühendsõnastikust[11].

[11] EKI Ühendsõnastik 2021: https://doi.org/10.15155/3-00-0000-0000-0000-08979L

Ühendsõnastiku sõnade loendi tegemiseks võtsin esmalt sõnastikust kõik sõnad, mis on potentsiaalselt spontaansust või mitteformaalsust väljendava märgendiga ehk sõnad, mis on märgitud kõnekeelseks, vulgaarseks või halvustavaks. Seejärel vaatasin saadud alamhulga käsitsi üle, et eemaldada ebasobilikke sõnu. Ebasobivateks pean mitmetähenduslikke sõnu, nagu *kapsas*, mis on saanud endale kõnekeelsuse märgenduse kasutamisest kulunud ja narmendava raamatu tähenduse tõttu, ning väga harva esinevaid ja vananenud sõnu, mis tänapäeva kontekstis leiduvad pigem vaid kirjanduses. Pärast nende eemaldamist jäi alles kaks loendit: 2972 sõna pikk üksiksõnade loend ja 468-pikkune mitmesõnaliste fraaside loend. See tõstis leksikonide arvu kaheteistkümnelt neljateistkümnele.

3.2.4. Leksikonide teisendamine tunnuseks

Otsustasin tööst välja jätta Ühendsõnastikus leidunud mitmesõnalised fraasid, kuna need nõuavad põhjalikku süntaksi analüüsi, mis kahjuks ei mahtunud magistritöö raamidesse. Võttes näiteks kõnekeelse väljendi *üles haipima*, saab kergesti üles leida laused, kus väljendi osad on kõrvuti, nagu “Toode on üles haibitud” ja “Ta haipis toodet üles”. Probleem seisneb aga lausetes, kus fraasi moodustavad sõnad ei ole kõrvuti. Näiteks “Ta haibib jälle üht toodet üles.” ja “Ta leidis haibitud toote poest üles.”, kus esimeses lauses tekib liitverb, aga teises lauses seda ei teki. Seepärast olekski selle ülesande jaoks vaja süntaksi analüüsi, et tuvastada lauses seotud üksuseid, mis ei asu teineteise kõrval. Probleemseks jäävad siiski aga laused nagu “Ta läks trepist üles haipima oma uut toodet.”, kus inimene saab selle sisust aru, kuid arvutile jääb segaseks, kas seoses on “trepist üles” või “üles haipima”.

Lisaks jätsin välja ka mõned enda koostatud tundmatuks jäänud sõnade loendid. Formaadivead jätsin välja, kuna need ei väljenda ei taotluslikku märgiasendust ehk tegu ei ole tahtliku otsusega väljendada stiili, ega ka spontaansusest, kiiresti kirjutamisest ja kontrollimata jätmisest tingitud kirjavigu. Jätsin välja ka numbrivead, sest vaadates allikaid, kus sellised sõnad esinesid, ei olnud tegu tekstidega, mis oleksid selgelt spontaansed või mitteformaalsed, nagu retseptid. Viimaks otsustasin spontaansuse ja formaalsuse hindamiseks välja jätta ka murdesõnad, kuna suur hulk nendest pärinesid

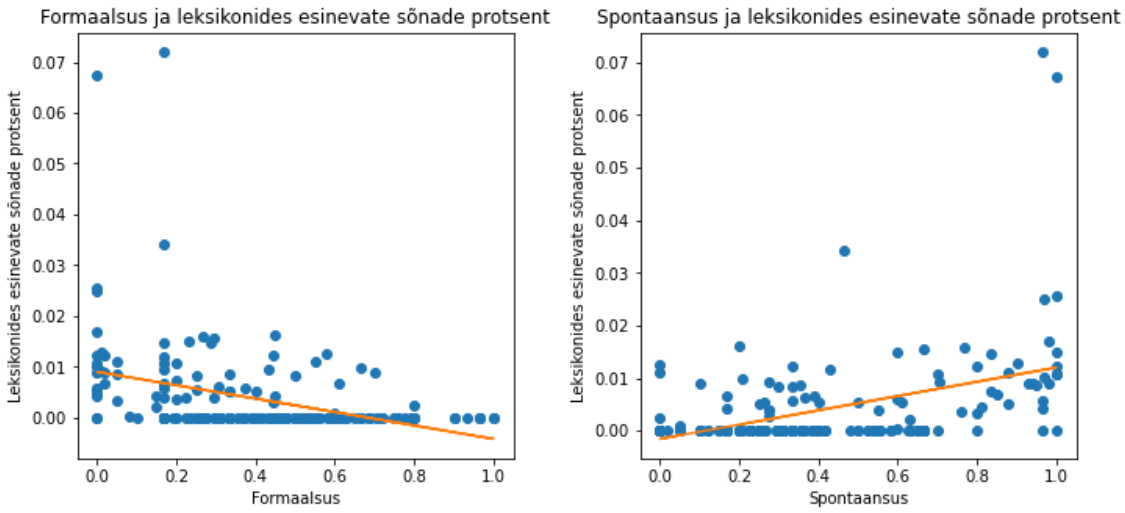
arhiividest ning arhiivilaadsetest blogidest, kus kogu tekst oli murdeline ning seega ei olnud tegu spontaansust või mitteformaalsust väljendavate sõnadega.

Lõplikus koodis otsustasin kõik kasutatavad leksikonid kombineerida üheks ühisleksikoniks. Individuaalsed alamleksikonid, kuigi potentsiaalselt kasulikud võimaldades anda eri kaale erinevaid kategooriad väljendavatele sõnatüüpidele, ei sisalda piisavalt sõnavara, mis esineks korpuses väga sageli. EstNLTKs tundmatu sõna märgendi saanud sõnade loendi alusel koostatud alamhulkadest leidus 100 000 teksti suuruses valimis näiteks ainult 8372 keelevea alamhulgas leiduva sõnaga teksti ehk ainult 8,4% tekstidest. Tasuks täheldada, et keelevigu esineb nimetatud loendite hulgast kõige enam. Teised alamhulgad olid esindatud vahemikus 3,5% (partiklite alamhulk 3466 tekstiga) kuni 1,1% (formaadivigade alamhulk 1058 tekstiga). Ühendsõnastiku alusel koostatud loendis esinevate sõnadega tekste leidus seevastu 32 474 ehk 32,5% valimis olevatest tekstidest sisaldasid vähemalt üht selles loendis olevat sõna.

Dimensioonide skoore mõjutab lõplikus leksikonis olevate sõnade protsent kõikidest teksti sõnadest. Vaatan iga sõne kohta, et kas vaadeldav sõne ise või selle lemma on leksikonides. Teen seda, kuna leksikonid koosnevad nii sõnavormidest kui ka sõnade algvormidest ning ainult ühte kasutades ei oleks tulemus õige.

Tunnus mõjutab nii formaalsust kui ka spontaansust. Formaalsust mõjutab leksikonides esinevate sõnade arv negatiivselt ja spontaansust positiivselt. Mida suurem on protsent, seda tugevam on mõju. Vahemikud, kui tugevalt mõjutavad tunnuse väärtused dimensioonide skoore, on lisades tabelis 15.

Kontrollisin tunnuse mõju suunda kummagi dimensiooni jaoks hinnangukorpuse abil. Seda näitab joonis 16, kus on kujutatud formaalsuse ja spontaansuse dimensioonide suhted leksikonis esinevate sõnade protsendiga. On näha, et formaalsuse suhe tunnusega on negatiivne ja spontaansuse suhe positiivne.



Joonis 16. **Formaalsuse ja spontaansuse dimensioonide suhted leksikonides esinevate sõnade protsendiga.**

4. Tunnuste kaalud

On intuitiivselt arusaadav, et hindamisel on mõned aspektid olulisemad kui teised. Näiteks valides kastist õuna, on mädaplekkide puudumine olulisem kui õuna punasus, kuid mõlemad on siiski olulised faktorid. Eeldan, et sarnane olukord on spontaansuse ja formaalsuse hindamisel, kus mõned tunnused on dimensiooni skoori hindamisel määravamad.

Bakalaureusetöös loodud mudelis olid aga kõik tunnused võrdsed ehk õuna hinnates oli mädaplekkide olemasolu sama oluline kui selle punasus. Mudelit koostades keskendusin tunnuste valimisele ja nende dimensioonide skoorideks teisendamisele ning seega soovisin mudeli edasi arendamiseks uurida, millisel määral tunnused dimensioonide skoore mõjutatavad ja seega millised on iga tunnuse tähtsused ehk kaalud.

Kaalude leidmiseks kasutasin andmestikuna hinnangukorpust ning nende arvutamiseks kasutasin programmeerimiskeelt R ja selle juhumetsade paketti randomForest[12]. Kasutasin juhumetsi, kuna tegu on algoritmiga, mis varieerib, milliseid tunnuseid igas alampuus hindamiseks kasutada ja seega võimaldab vaadata, millised tunnused aitavad

[12] R-i pakett randomForest: https://cran.r-project.org/web/packages/randomForest/index.html

kõige paremaid skoore kätte saada. Kuna algoritm kasutab tunnuste kombinatsioone ja mitte ainult üksikuid tunnuseid, on võimalik, et mõni tunnus, mille puhul on regressioonijoon dimensiooni ja tunnuse vahel väikese suunaga, nagu joonisel 10 kujutatud formaalsuse ja *nud*-partitsiibi vormis verbide protsendi vahelise suhte puhul, on olulisem, kui vaid regressioonijoone tõus mõista annaks.

Tegin kaks juhumetsa, üks spontaansuse ja teine formaalsuse hindamiseks. Spontaansuse hindamise juhumets võtab ennustatavaks tunnuseks korpuses oleva spontaansuse skoori ning tunnusteks varasemalt määratud spontaansuse tunnused. Juhumets selgitas 56,85% spontaansuse hinnangu varieerumisest. Formaalsuse hindamise juhumets võtab ennustatavaks tunnuseks korpuses oleva formaalsuse skoori ning tunnusteks varasemalt määratud formaalsuse tunnused. Juhumets selgitas 63,89% formaalsuse hinnangu varieerumisest.

Seejärel vaatasin juhumetsa seest tunnuste olulisusi. Kummagi dimensiooni puhul vaatasin iga tunnuse %IncMSE-d (*keskmise ruutvea tõus*), mis väljendab, kui palju tõuseb juhumetsa keskmine ruutviga (MSE ehk *mean squared error*), kui muuta tunnuse väärtust. Seega, mida kõrgem on keskmise ruutvea tõus, seda olulisem on tunnus. %IncMSE annab ligikaudse arvutuse dimensiooni iga tunnuse olulisusele, mida kasutasin edasi dimensioonide hindamisel kaaludena. Iga tunnuse puhul korrutasin selle esialgse arvulise tulemuse läbi vastava %IncMSE-ga ning saadud korrutise liitsin dimensiooni hinnangule. Kuna kõige suurema ja kõige väiksema %IncMSE vahel on väga suur vahe, otsustasin neid edasi vaadelda ja katsetada, kuidas nende lähendamine mõjutab dimensioonide skoore.

Formaalsuse puhul oli kõige olulisem tunnus käändsõnade protsent tekstis kaaluga 20,75 ja kõige vähem olulisem tunnus kaudse kõneviisi protsent kaaluga 1,96. Spontaansuse puhul oli samuti kõige olulisem tunnus käändsõnade protsent tekstis kaaluga 21,00, aga kõige vähem olulisem tunnus oli tajuverbide protsent kaaluga 0,03.

Spontaansuse puhul oli kõige olulisema tunnuse kaal 700 korda suurem kõige vähem olulise tunnuse kaalust, seega katsetasin ka kaalude tasandamist. Arvestades, et neutraalne kaal on üks ja mitte null, sest see annaks läbi korrutades alati nulli, otsustasin

katsetada viia kaalusid ühele lähemale, kasutades juurimist. Võtsin katsetamiseks juurijaks arvu 1,4, sest see on tasakaal kaalude ühele lähemale viimise ja kaalude väärtuste eristamise vahel. Kõrgemad juurijad viisid kaalude väärtused teineteisele liiga lähedale ehk kaalud hakkasid teineteisega väga sarnanema, mis oleks teinud kaalude kasutamise ebavajalikuks. Väiksemad juurijad aga ei tasandanud kaale piisavalt, jättes kõige kõrgema ja kõige madalama kaalu vahele näiliselt liiga suure vahe. Tegu on siiski ainult katselise juurijaga ja selle ülesanne on aidata vaadata, kas %IncMSE kaudu saadud kaalude tasandamine võiks olla vajalik.

Kaalude katsetamiseks arvutasin kummagi dimensiooni jaoks kolm skoori: kaaludeta skoor ehk skoor, kus kõik tunnused on võrdse kaaluga; tasandamata kaaludega skoor, kus kaaluks on tunnuse %IncMSE, ja tasandatud kaaludega skoor, kus %IncMSE on juuritud 1,4-ga. Lisaks teisendasin kõik skoorid vahemikku 0 kuni 1, kus null tähistab kõige väiksemat skoori enne teisendamist ja üks kõige suuremat skoori, mis viib skoorid samasugusesse vahemikku, nagu need on hinnangutega korpustes.

Tegin seda kasutades lineaarteisendust, mis on kujutatud võrrandil 1. *x* tähistab teisendamata hinnangut ning *y* teisendatud hinnangut. a_1 tähistab teisendamiseelset miinimumväärtust, antud juhul madalaimat teisendamiseelset skoori, ning b_1 tähistab teisendamiseelset maksimumväärtust, antud juhul suurimat teisendamiseelset skoori. a_2 ja b_2 tähistavad teisendamisjärgseid miinimum- ja maksimumväärtuseid, antud juhul vastavalt 0 ja 1.

$$y = \frac{(x - a_1) \times (b_2 - a_2)}{(b_1 - a_1)} + a_2$$

Võrrand 1. **Lineaarteisenduse valem.**

Kuna teisendamisjärgsed miinimum- ja maksimumväärtused on püsivalt 0 ja 1, saab valemit lihtsustada, mis on kujutatud võrrandil 2.

$$y = \frac{(x - a_1)}{(b_1 - a_1)}$$

Võrrand 2. **Lihtsustatud lineaarteisenduse valem.**

Seejärel vaatasin, kuidas seostuvad omavahel dimensioonide skoorid ja inimhinnangud. Lahutasin igalt skoorilt vastava dimensiooni hinnangu ja võtsin selle absoluutväärtuse. Tegin seda kõigi 180 tekstiga ning arvutasin kummagi dimensiooni jaoks iga kolme skoori (kaaludeta, tasandamata kaaludega ja tasandatud kaaludega mudelite) jaoks keskmise absoluutvea (MAE ehk *mean absolute error*), et vaadata, millised kaalud on kõige lähemal inimhinnangule.

Formaalsuse puhul osutus kõige kehvemaks tasandatud kaaludega skoor, mille MAE oli 0,1945. Edasi tuli ilma kaaludeta skoor, MAE-ga 0,1903 ja kõige parem oli tasandamata kaaludega skoor, mille MAE oli 0,1889. Spontaansuse puhul oli olukord teistsugune, sest selle puhul osutus kõige kehvemaks ilma kaaludeta skoor, mille MAE oli 0,1731. Seejärel tuli tasandamata kaaludega skoor, MAE-ga 0,1701, ja parimaks osutus tasandatud kaaludega skoor, mille MAE oli 0,1654. Lõplikus mudelis otsustasin seega kasutada formaalsuse hindamiseks tasandamata kaale ja spontaansuse hindamiseks tasandatud kaale.

Olen lisanud kaks tabelit, mis kujutavad nii tasandatud kui ka tasandamata kaaludega mudelite kaale. Tabel 1 näitab iga tunnuse kaale formaalsuse hindamisel ning tabel 2 näitab spontaansuse mudeli kaale. Mõlema tabeli puhul on tasandamata kaal sama, mis juhumetsa %IncMSE, ümardatud kahe komakohani, ning tasandatud kaal selle 1,4-ga juurimise tulemus.

Tabel 1. **Formaalsuse kaalude tabel.**

Tunnuse nimi	Tasandamata kaal	Tasandatud kaal (juurija 1,4)
Käändsõnade protsent	20,75	8,72
Keskmine lemmapikkus	18,50	8,04
Verbide esimese isiku protsent	12,82	6,19
Verbide kolmanda isiku protsent	12,72	6,15
Umbisikulise tegumoe protsent	9,83	5,12
Emotikonide arv	9,64	5,05
Sõnavara mitmekesisus	8,80	4,73
Asesõnade esimese isiku protsent	7,97	4,40
Leksikonides esinevate sõnade protsent	6,30	3,72
Verbide teise isiku protsent	5,48	3,37
Asesõnade teise isiku protsent	4,19	2,78
nud-partitsiibi vormis verbide protsent	4,03	2,71
Asesõnade kolmanda isiku protsent	2,44	1,89
Kaudse kõneviisi protsent	1,96	1,62

Tabel 2. **Spontaansuse kaalude tabel.**

Tunnuse nimi	Tasandamata kaal	Tasandatud kaal (juurija 1,4)
Käändsõnade protsent	23,01	9,39
Puuduva suure algustähega sõnade protsent	16,53	7,42
Keskmine lemmapikkus	14,95	6,90
Puuduva tühikuga kirjavahemärkide arv	11,40	5,57
Emotikonide arv	11,15	5,60
Asesõnade esimese isiku protsent	9,86	5,13
Leksikonides esinevate sõnade protsent	9,03	4,82
Sõnavara mitmekesisus	8,29	4,53
Läbinisti suurtäheliste sõnade protsent	6,11	3,64
Sõnasiseste korduste arv	3,14	2,26

5. Tulemused

Et analüüsida, kas ja kuidas on mudeli edasi arendamise eesmärgil sisse viidud lisad ja muutused mõjutanud mudeli hinnanguid, võrdlen edasiarenduse hinnanguid algse, bakalaureusetöö (Gailit 2021) raames loodud mudeli hinnangutega. Teen seda kahel viisil, esiteks kasutades hinnangukorpust, mida olen kasutanud kaalude arvutamiseks, ning teiseks võrreldes kümne teksti algse ja arendatud mudeli hinnanguid. Viimane võrdlus on oluline, et saaks vaadata, kuidas arendatud mudel hindab uusi tekste, mida ei ole kasutatud mudeli arendamiseks.

5.1. Hinnangukorpuse võrdlus

Mudelite võrdlemiseks kasutades hinnangukorpust leian iga selle teksti jaoks kaks veamäära: bakalaureusetöö mudeli veamäära ja edasiarendatud mudeli veamäära. Veamäärad arvutan võttes mudeli hinnangu ja inimhinnangu vahe absoluutmäära, ehk vaatan, kui palju erineb mudeli hinnang inimeste hinnangust. Seejärel arvutan veamäärade keskmised, liites iga teksti veamäärad kokku ja jagan kõikide tekstide arvuga. Teen seda iga nelja hinnangu jaoks – bakalaureusetöö formaalsuse ja spontaansuse hinnangud ning edasiarendatud mudeli formaalsuse ja spontaansuse hinnangud. Saadud arvud väljendavad, kui palju erinevad keskmiselt mudelite hinnangud inimhinnangutest ning mida väiksem on veamäär, seda sarnasem on mudeli hinnang inimhinnangule ja seega seda parem on mudel vaadeldava andmestiku hindamiseks.

Ümardatuna nelja komakohani, on bakalaureusetöö mudeli formaalsuse keskmine veamäär 0,29 ja spontaansuse keskmine veamäär 0,37. Kuna hinnangud on skaalal null kuni üks, on tegu suure veamääraga, mis tõestab, et mudeli edasi arendamine oli vajalik. Edasi arendatud mudeli formaalsuse keskmine veamäär ümardatud kuni kahe komakohani on 0,19, mis on 0,1 võrra ehk 35,48% parem, ning spontaansuse keskmine veamäär on 0,17, mis on 0,21 võrra ehk 55,61% parem. Mõlema dimensiooni puhul on veamäär paranenud väga suurel määral, mis tähendab, et sisse toodud parandused ja täiendused on täitnud oma eesmärgi, tehes mudelit paremaks.

Tuleb aga täheldada, et tegu ei ole ideaalse võrdlusega. Esiteks kasutasin ma bakalaureusetöös hinnangute skaalana vahemikku -1 kuni 1, kus -1 väljendas ebadimensionaalsust, 1 dimensionaalsust ja 0 neutraalsust. Mudeli edasiarenduses olen aga ebadimensionaalsuse välja jätnud ehk skaala väljendab dimensiooni esinemise tugevuse määra – 0 tähistab, et dimensiooni tekstis ei leidu, ja 1 tähistab, et dimensiooni leidub tekstis väga tugevalt. Kuigi hinnangud ei väljenda täpselt sama ideed, on need siiski võrreldavad ning nende võrdlemine annab ülevaate, kuidas mudeli arendused on hinnanguid muutnud. Olen teisendanud bakalaureusetöö hinnangu vahemikku 0 kuni 1, liites algsele väärtusele ühe ja jagades saadud summa kahega.

Lisaks kasutan hindamiseks sama andmestikku, mille põhjal olen arvutanud tunnuste kaalud. See tähendab, et edasi arendatud mudeli keskmine veamäär võib olla parem ainult seepärast, et see on seadistatud just seda kindlat andmestikku hindama, ning mõnda varem vaatlemata andmestikku kasutades võivad tulemused olla palju halvemad.

5.2. Tekstide analüüsimine

Et saada paremat ülevaadet, kuidas hinnangud erinevad algse ja arendatud mudelite vahel, olen võtnud Ühendkorpuse valimist kümme suvalist lühikest teksti, mida ei ole varem hinnanud ega kasutanud kaalude arvutamiseks. Võrdlemise selgemaks tegemiseks olen ka siin teisendanud bakalaureusetöö mudeli hinnangud vahemikku null kuni üks ehk 0,5 tähistab selle mudeli puhul dimensiooni suhtes neutraalset teksti, arendatud mudeli puhul aga vähesel või mõõdukal määral dimensiooni sisaldavat teksti. Edasi olen välja toonud kõik kümme teksti ning võrrelnud nendele antud algse ja arendatud mudelite hinnanguid kuni kahe komakohani ümardatuna.

Tekst 1 on Haldja Hambaravi veebilehekülje privaatsuspoliitika seisuga 11. detsember 2019. Lehekülg kirjeldab, kuidas ja mis eesmärkidel lehekülje kasutajate andmeid kogutakse:

(1) *Privaatsuspoliitika*

Andmete töötlemise põhimõtted

Teie isikuandmete vastutustundlik kaitsmine on meie jaoks oluline. Isikuandmete töötlemise poliitikas kirjeldatakse, milliseid andmeid ja millisel eesmärgil kogutakse. Andmeid töödeldakse kooskõlas kehtivate seadustega ning isikuandmete töötlemisel lähtume alati Teie huvidest, õigustest ja vabadustest. Isikuandmete töötlemine toimub viisil, mis tagab isikuandmete asjakohase turvalisuse, sealhulgas kaitseb loata või ebaseadusliku töötlemise eest ning juhusliku kaotamise, hävitamise või kahjustumise eest, kasutades mõistlikke tehnilisi või korralduslikke meetmeid. Kliendi isikuandmete vastutav töötleja on Haldja Hambaravi OÜ, registrikoodiga 11564177, aadressiga Pärnu mnt 30, Märjamaa alev, Märjamaa vald, Raplamaa 78301.

Isikuandmete töötlemise eesmärgid

Isikuandmete töötlemise eesmärgiks on veebilehe poolt kogutavate isikuandmete analüüsimine ning läbi selle ettevõtte poolt pakutavate teenuste kasutamise lihtsustamine. Lisaks aitab andmete kogumine pakkuda kasutajatele huvipakkuvat veebilehe sisu.

Veebilehe külastaja õigused isikuandmete töötlemisel

Veebilehe külastaja omab seoses oma isikuandmete kasutamisega seadusest tulenevaid õigusi:

õigus loobuda uudiskirjadest/personaalsetest pakkumistest.

Veebilehel kogutavad andmed

Seoses sellega, et veebilehel on kasutusel Google Analytics statistika liides, kogub veebileht järgnevaid isikuandmeid:

Lehel registreerudes küsitakse ja/või hangitakse läbi äriregistri (juhul, kui tegu on ettevõttega) järgnevaid andmeid:

(Ühendkorpus 2019 hambahaldjas.ee, doc id = 5547770)

Algne mudel on andnud tekstile formaalsuse skooriks 0,72 ja spontaansuse skooriks 0,53 ning arendatud mudel annab formaalsuse skooriks 0,81 ja spontaansuse skooriks 0,2. Selle teksti puhul on arendatud mudeli hinnangud paranenud võrreldes algse mudeliga. Formaalsuse skoori suurenemine on hea õigusteksti kui stereotüüpilise formaalse teksti puhul. Kuigi hinnang võiks kõrgem olla, ei saa see aga tekstist lõikude puudumise tõttu maksimaalne ehk üks olla. Kuna õigustekst on ka tüüpiline mittespontaanne tekst, on selle skoori vähenemine arendatud mudeli puhul hea, sest see tähistab, et tekstis leidub spontaansuse dimensiooni vähe, mis on teksti puhul tõene.

Tekst 2 on veebipoe Liann-Lõngad kasutustingimused:

> (2) *Kasutustingimused*
>
> *Tere tulemast Liann-Lõngad veebilehekülje! Meil on hea meel, et tunnete huvi meie toodete vastu, samas Teie privaatsus on meile väga tähtis. Liann-Lõngad OÜ kasutab teie isikuandmeid üksnes selleks, et pakkuda teile parimat teenindust, töödelda teie tellimust ja tarnida teile tellitud kaup ning hallata teie kontot.*
>
> *Privaatsuspõhimõtted*
>
> *Kui külastate meie veebisaiti, siis üldjuhul teie isikuandmeid ei salvestata. Teie kui internetikasutaja jääte anonüümseks, sest me analüüsime seda teavet vaid statistilisel eesmärgil (nt külastuste arv lehe kohta). Meie veebisaidi külastajad jäävad anonüümseteks internetikasutajateks. Teie isikuandmed saadakse üksnes siis, kui annate need vabatahtlikult, näiteks saates läbi veebi teabepäring või registreerudes püsikliendiks ja tehes ostutellimus. Meie veebisaidi kaudu edastatud andmed krüpteeritakse SSL/TLS krüpteeringuga HTTPSi turvaprotokolli kasutades.*
>
> *Liann-Lõngad OÜ ei edasta andmeid kolmandatele isikutele, välja arvatud üksnes sellised partnerid, kes peavad ellu viima ostuprotsessi, mille klient tellimuse esitamisega algatas (pangamaksed läbi epoe ja transport kliendi valitud kanalit pidi). Kui otsustate kasutada maksemeetodina PayPali platvormi, siis teavitab meie veebisait automaatselt PayPali rakendust teie arvega seotud andmetest*

(nimi ja aadress) ja kontaktandmetest (telefon ja e-posti aadress). Te saate selle teabe esitamisest loobuda, kui tühistate PayPali kaudu maksmise soovi. Kui otsustate siiski teha makse PayPali kaudu, siis kohustute järgima sel hetkel kehtivaid PayPali privaatsuspõhimõtteid.

Kaupade ostmine e-poest

(Ühendkorpus 2019 www.liann.ee, doc id = 5555527)

Algne mudel on andnud formaalsuse skooriks 0,49 ja spontaansuse skooriks 0,58 ning arendatud mudel annab formaalsuse skooriks 0,56 ja spontaansuse skooriks 0,24. Nagu tekst 1 puhul, ei saa olla kindel, millest see eripära on tekkinud. Ka sisu poolest meenutab vaadeldav tekst teksti 1, olles privaatsuspoliitika, kuid tekstide kirjutamise stiil on erinev. Kui tekst 1 oli vormistuselt sarnane õigustekstidega, on tekst 2 palju pikemate lõikudega ning sisaldab keelekasutust, mille eesmärk on tekitada familiaarsust, nagu kohe teksti alguses lugeja tervitamine. Seega võib öelda, teksti formaalsuse hinnangu suurenemine vähesel määral ebaformaalsest hinnangust mõõdukalt formaalsust sisaldavaks on mõistlik muutus. Võrreldes teksti 2 tekstiga 1, saab öelda, et tekst 2 sisaldab vähem formaalsust kui tekst 1, mis tuleb välja ka nende hinnangutest. Spontaansuse langemine on samuti väga hea, kus kuigi teksti vormistus ei ole tekstitüübile tavaline, on tegu siiski privaatsuspoliitikaga, mis tingib selle läbimõeldud kirjutamise, viidates asjaolule, et tekst ei saa olla kiiresti kirjutatud ega ka vähemalt autori poolt toimetamata.

Tekst 3 pärineb Statistikaameti Väliskaubanduse rakenduse veebileheküljelt ning see kirjeldab lühidalt, mis riikidest imporditi Eestisse ning mis riikidesse eksporditi Eestist kroomoksiide ja -hüdroksiide aastal 2018:

(3) *Infograafikud*

Kroomoksiidid ja -hüdroksiidid on kaubavahetuse väärtuse poolest 575. toode HS4 tasemel.

2018 olid peamised riigid, kuhu toodet Kroomoksiidid ja -hüdroksiidid eksporditi, Saksamaa (615 tuhat eurot), Itaalia (291 tuhat eurot), Hispaania (81,3 tuhat eurot), Iraan (34,9 tuhat eurot) ja Poola (34,2 tuhat eurot).

Samal perioodil olid peamised riigid, kust toodet Kroomoksiidid ja -hüdroksiidid imporditi, Kasahstan (883 tuhat eurot), Saksamaa (20,7 tuhat eurot) ja Holland (3,73 tuhat eurot).

Kroomoksiidid ja -hüdroksiidid on 4-kohalise koodiga toode (HS4 ID 2819).

(Ühendkorpus 2019 data.stat.ee, doc id = 5016593)

Bakalaureusetöös loodud algne mudel on andnud formaalsuse skooriks 0,52 ja spontaansuse skooriks 0,59 ning arendatud mudel annab formaalsuse skooriks 0,57 ja spontaansuse skooriks 0,34. Arendatud mudeli spontaansuse skoor on madalam ning seepärast algsest mudelist parem, sest teksti informatsioonitihedus ei ole omane kõnekeelele ega ka kiirele kirjutamisele ning seega ei esine tekstis spontaansust tugeval määral. On aga aru saada, et spontaansust selles kindlas tekstis mingil määral siiski esineb, mis väljendub pideva liigse suure algustähega sõna *kroomoksiidid* puhul. See võib aga ka tähistada, et tekst on genereeritud automaatselt, kus lüngad on tabeli abil automaatselt täidetud, kuid vormistus on unustatud õigeks määrata. Teksti formaalsuse skoor suurenes vähesel määral. Kuna tekst on ametlik infograafikute ülevaade, võib öelda, et formaalsuse dimensiooni esineb selles mõõdukal kui isegi mitte suuremal määral, ning seepärast on skoori suurenemine tähis, et arendatud mudel on vähesel määral paranenud.

Tekst 4 on ülevaade sümbolistlikust kunstist, peamiselt 19. sajandi lõpus. Tekstis on välja toodud mitmeid riike, nende sümboliste ja teoseid:

(4) *Prantsuse kunstnikel oli suhteliselt vähem kalduvust sümbolismile, hoopis enam levis see aga saksa kunstnike seas. Kõige huvitavamateks prantsuse sümbolistideks on Odilon Redon (1840-1916) oma väga fantaasiaküllaste nägemuspiltidega ning peamiselt seinamaalide autor Pierre Puvis de Chavannes*

(1824-1898). Mõlema teosed on vägagi meeldivad ka oma välise, puhtmaalilise külje poolest. Puvis de Chavannes näiteks eelistas tagasihoitud sinakashalle toone ning selgeid lihtsaid vorme.

Samuti Hollandist, kus Jan Toorop (1858-1928) lõi kuidagi haiglasliku ja mürgise meeleoluga pilte. Neil näeme külmi rohelisi ja siniseid toone ning venitatud valgenäolisi inimesi.

Väga tuntud on šveitslase Ferdinand Hodleri (1853-1918) müstilise alatooniga maalid.

Inimlikult mõistetavam oli norralane Edvard Munch (1863-1944) - tema maal "Elutants" näiteks väljendab selliseid igaühele arusaadavaid asju nagu noorus ja vanadus, lootused ja pettumused; kuupaistesel rannal tantsivate naiste puhasvalgete, lõõmavpunaste ja mustade rõivastega annab ta edasi inimelu kulgu.

Väga kaunid, erootilised ja müstilisest hingusest kantud on austerlase Gustav Klimti (1862-1918) teose

19. sajandi lõpu kirjut kunstipilti teeb veelgi keerukamaks see, et tihti on tolleaegsed erinevad kunstisuunad omavahel segunenud. Näiteks võisid nii postimpressionistid kui sümbolistid käsutada juugendlikke väljendusvahendeid. Samal ajal õhkus teinekord postimpressionistide loomingust sümbolistlikku salapära.

Põhjamaades, Venemaal ja ka Eestis, arenes sümbolism nn rahvusromantiliseks suunaks. Püüti kajastada oma maade kauget sangarlikku minevikku, elustada rahvaluule kaunis maailm, kujutada uudselt armastatud kodumaastikke. Viimased on Põhjamaade tolleaegsetel maalidel enamasti dekoratiivsed, rasketes oranžides sügisevärvides. Põhjamaade kunstnikest on väga kuulus Kaseli Gallén-Kallela (1865 – 1931) oma soome rahvaeepose "Kalevala" ainetel loodud teostega.

(Ühendkorpus 2019 hingetugi.onepagefree.com, doc id = 5478201)

Algne mudel on andnud formaalsuse skooriks 0,68 ja spontaansuse skooriks 0,5 ning arendatud mudel annab formaalsuse skooriks 0,78 ja spontaansuse skooriks 0,17. Teksti spontaansuse skoori suur vähenemine arendatud mudeli puhul on hea, sest tekst on stiili ja informatsioonitiheduse poolest sarnane õpikute tekstidega ning võib eeldada, et autor on enne avaldamist teksti pikemat aega kirjutanud ja toimetanud. Samadel põhjustel võib ka formaalsuse hinnang olla suurenenud. Kuna tekst on kirjutatud pigem ilmekalt, kirjeldades maale loominguliselt kasutades fraase nagu "*haiglasliku ja mürgise meeleoluga*", ei saa öelda, et tekst on tugevalt formaalne, eriti võrreldes teksti teadusartikli või ka lausa Wikipedia artikliga sümbolismi kohta. Kuid tõus on väike ning leian, et see on mõlemal mudelil teksti jaoks adekvaatne formaalsuse hinnang.

Tekst 5 on 2016. aasta juunis avaldatud postitus Jaan Kaplinski blogist, mis on väga üldistatult kirjutatud rahvusvahelise poliitika ja sõja teemadel:

> (5) *Mul oli väga hää meel lugeda Saksamaa välisministri Frank-Walter Steinmeieri sõnavõttu, kus ta arvas NATO lakkamatutest manöövritest siin kandis sama, mis mina. Steinmeieri kõige karmim ütlus oli: "Was wir jetzt nicht tun sollten, ist durch lautes Säbelrasseln und Kriegsgeheul die Lage weiter anzuheizen" -- See, mida me praegu mitte tegema ei peaks, on valju mõõgatäristamise ja sõjakisaga olukorda veel enam üles kütta. Nii ongi. Meil oskas spordiajakirjanik Roosna oma arvamuse naabermaast võtta kokku leheloo päälkirjas "Venemaa mõistab ainult ühte keelt. " Vägev üldistus. Ühe päälkirjaga Venemaa paika pandud. Ainult et vägisi kipub tulema mõte, et kui Saksa diplomaatia juht mõtleb pääga, siis mõni ajakirjanik ühe teise kohaga, mida ta ka mainimata ei jäta, sobigu see või ei. Muidugi võeti ja võetakse Steinmeieri sõnad meil, kus mõõgatäristamine ja unistus Venemaale jalaga... on saanud meie hommiku- ja õhtupalve aseaineks, vastu ulgumise ja hammaste kiristamisega. Aga ehk on tervel mõistuselgi Ida-Euroopas veel paigake ja võimalus end kuuldavale tuua. Ja panna inimesed mõtlema sellele, et Eesti ja Eesti rahvas ei elaks üle veel üht sõda, üht ränka konflikti. Ja mõõgatäristamine ja sõjakisa ei pruugi sõjavõimalust meist eemal hoida, vaid tuua meid veel üheks ohvriks Ameerika geopoliitilises mängus. Oma iseseisvuse oleme juba käest andnud, vähemalt*

alates Iraagi sõja eelmängust 2003, kas järgmisena oleme valmis ära andma oma elu ühe ookeanitaguse suurriigi huvide nimel?

(Ühendkorpus 2019 jaankaplinski.blogspot.com, doc id = 3734073)

Algne mudel on andnud formaalsuse skooriks 0,4 ja spontaansuse skooriks 0,69 ning arendatud mudel annab formaalsuse skooriks 0,44 ja spontaansuse skooriks 0,43. Spontaansuse dimensiooni tekstis esinemise vähenemine on mõistlik, kuna tekst sarnaneb stiili poolest rohkem ajakirjanduslikele arvamuslugudele kui tüüpilistele blogipostitustele, nagu tekstid 8, 9 ja 10. Võib eeldada, et algse mudeli kõrge spontaansuse skoor on tingitud tundmatuks jäänud sõnade tõttu, sealhulgas tekstis esinev saksakeelne tsitaat, ja diftongi *ea* asendamisel *ää*-ga, mille EstNLTK võib olla märkinud kirjavigaseks, kuigi tegu on pigem murdelisusega. Formaalsuse hinnangu tõusmine on samuti positiivne. Kuid tekst on kirjutatud ilmekas stiilis, nagu tekst 4, ning sisu poolest on tekst arutlev, mistõttu ei saa teksti pidada tugevalt formaalseks. Seepärast leian, et tekstis esineb dimensiooni mõõdukalt, mida ka arendatud mudel väidab, kuigi ka veidikene kõrgemad hinnangud ei oleks ebasobivad.

Tekst 6 on informatsioonileht Eesti Vabariigi 100. sünnipäeva tähistava kanuumatka kohta. Tekst on pigem hübriidne, sisaldades nii kirjeldavaid lõike kui ka matka informatsiooni ja registreerimise tingimusi:

(6) *Ühine kanuuretk Põltsamaa jõel*

Hea matkahuviline, ühine meiega ja tähistame koos juubeliaastat!

Ootame 11. augustil algusega kell 12 toimuvale juubelimatkale kokku vähemalt 100 matkasõpra, et pakkuda koosolemise rõõmu, tähistada Eesti Vabariigi juubeliaastat ning luua ühiseid ja kauakestvaid mälestusi.

Kanuumatk on hea võimalus korraks aeg maha võtta, igapäevamürast eemale saada ning kaaslastega koos imelist loodust nautida. Põltsamaa jõgi on ainuke Eestis, mis voolab nelja maakonna piires. See läbib ka kunagist Eestimaa ja Liivimaa piiri, olles oluline lüli mööda veeteid liikudes.

Oluline info:

Matka pikkus: 12 km

Matka marsruut: Rutikvere – Pajusi

Stardipunkt on avatud kell 12-15 (matkajad saavad ise sobivaima stardiaja valida)

Matka kestus kokku: 2-3h (olenevalt valitud tempost)

Osalustasu: 20 eurot/inimene; kuni 10-aastased (k.a) lapsed 15 eurot/inimene

Osalustasu sisaldab matkale minekuks vajalikku varustust (päästevest, mõla, koht kanuus), instruktaaži ja matkajuhi teenust.

Matkal osalemiseks on vajalik registreeruda!

See kanuuretk ei ole mõeldud võidusõitmiseks vaid rahulikuks kulgemiseks. Nii jääb vaikselt allavoolu liikudes aega ka looduseilu nautida, uudistada, mida koprad on korda saatnud jm põnevat. Kuna Põltsamaa jõgi on üsna rahuliku iseloomuga, sobib matk nii noortele, lastega peredele kui ka juba kogenud matkajatele. Matka lõpetame üheskoos muljeid vahetades ja sooja teed rüübates.

peakate

Registreerumise tingimused

Kanuuretkele registreerumiseks saata info osalejate arvu ja nimedega e-mailile info@kanuumatkad.ee või helistada telefonil +372 518 9322.

2.1 Registreerudes rohkem kui 28 päeva enne kanuuretke toimumist, tuleb osalustasu tasuda pangaülekandega 14 päeva jooksul pärast registreerumist ja arve saamist 2.2 Registreerudes 28 kuni 14 päeva enne kanuuretke toimumist, tuleb osalustasu tasuda pangaülekandega 7 päeva jooksul pärast registreerumist

ja arve saamist 2.3 Registreerudes 14 ja vähem päeva enne kanuuretke toimumist tuleb osalustasu tasuda kohe pärast registreerumist ja arve saamist.

Kanuuretkel osalemise tühistamisel lähtutakse järgmistest korraldaja poolt kehtestatud tingimustest:

3.1 Korraldajal on õigus tühistada kanuuretkele registreerumine, mille eest ei ole tähtajaks tasutud

3.2 Osaluse tühistamisel peab korraldaja osalustasust kinni teenustasu

(Ühendkorpus 2019 kanuumatkad.ee, doc id = 4782214)

Algne mudel on andnud formaalsuse skooriks 0,54 ja spontaansuse skooriks 0,68 ning arendatud mudel annab formaalsuse skooriks 0,61 ja spontaansuse skooriks 0,41. Spontaansuse hinnangu vähenemine on väga hea, kuna algse mudeli mõõdukalt positiivne skoor on ebaadekvaatne. Kuigi võiks öelda, et kanuuretke kirjeldavad lõigud on pigem spontaansemad, on hübriidsel tekstil lisaks õigustekstidele sarnane tingimuste loend, mille ebaspontaansus tingib teksti vähese kuni mõõduka spontaansuse hinnangu. Formaalsuse hinnangu suurenemine on samadel põhjustel positiivne, sest algse mudeli negatiivne formaalsuse hinnang ei ole kuidagi sobilik arvestades, et osa teksti on stiililt sarnane õigustekstidega.

Tekst 7 on ametüstist küünlaaluse tootekirjeldus esoteerilisest veebipoest, mis kirjeldab, millistel eesmärkidel ja millega koos toodet kasutada:

(7) *Geoodid on suurepärased aurapuhastajad. Aurakeha tervena ja leketeta hoidmine on vajalik teie spirituaalse ja emotsionaalse keha jaoks sama oluline nagu kehale on*

vajalik vesi. Sellest sõltub teie heaolu ja tervislik seisund!

Ametüst sobib asetamiseks igale poole. Lähtu oma sisemisest vajadusest ja oma soovidest intuitsiivselt. Ta on õiges kohas siis, kui oled selle sinna

intuitiivselt asetanud.

Jõukuse edendamiseks põleta alusel lillat tooni teeküünlaid ja Kaneeli viirukeid.

Tervise ja perekonna heaoluks põleta kollasied teeküünlaid ja Kummeli või Sandlipuu viirukeid.

(Ühendkorpus 2019 www.marilynkerro.ee, doc id = 3750123)

Algne mudel on andnud formaalsuse skooriks 0,5 ja spontaansuse skooriks 0,54 ning arendatud mudel annab formaalsuse skooriks 0,66 ja spontaansuse skooriks 0,2. Tootekirjeldused on üleüldiselt väga varieeruva formaalsusega ning vaadeldav tekst on pigem vähem kui rohkem formaalne. See asjaolu tuleb kõige paremini välja kolmandast lausest, mis sisaldab rõhutamise eesmärkidel hüüumärki. Seega on formaalsuse hinnangu suurenemine pigem negatiivne. Spontaansuse hinnangu vähenemine väga suurel määral on samuti ebasobiv, sest kuigi tüüpiline tootekirjeldus on tavaliselt toimetatud, on aru saada, et vaadeldav tekst seda ei ole. Toimetamise puudumist on näha peale formaadis lausesiseste reavahetuste ka paari kirjavea kaudu, kus sõnad *intuitsiivselt* ja *kollasied* sisaldavad näpuvigu. Need vead ei leidu leksikonides, kuid EstNLTK kirjavigade analüsaator on need aga tõenäoliselt vigadeks märkinud, mis on üheks põhjuseks, miks algne mudel on andnud tekstile õigema hinnangu. Spontaansuse hinnangu liigne väiksus tuleb eriti välja, kui võrrelda teksti varasemalt mainitud madala spontaansusega tekstidega, nagu tekst 4 või 2, kus viimane on saanud vaadeldavast tekstist kõrgema spontaansuse hinnangu, kuid inimene hindaks vastupidiselt, andes tekstile 7 kõrgema hinnangu kui tekstile 2.

Tekst 8 on blogipostitus 2019. aasta augustist, kus kirjeldatakse üht päeva pikemast reisist:

(8) *Kuna majakas on parasjagu ülerahvastunud ja meil palutakse oodata, teeme kerge näksi.*

"See batoon väga magus pole," pakub E K-le. "See on puuvillane," kinnitab H.

Pilet maksab 80 senti. No me saame aru, et omi hoitakse, aga koorigu vähemalt välismaalasi!

Poodi minemiseks on vaja liikuda 5 km sisemaale. Liigume. Korjame metsaaluse pohladest ja mustikatest lagedaks. Põldude vahel võtab K laulujoru üles: "Pūt, vējiņi, dzen laiviņu, Aizdzen mani Kurzemē. " Rohkem ei meenu. Küllap kaaslastele see isegi meeldib. Užava on kena väike puhas asula teeäärsete viljapuude, silla, laululava ja puhkekohtadega.

Tänu sillale saame oluliselt lõigata.

E ja H lähevad poodi ja K jääb kotte valvama. Vastassuunast saabub grupp rootslasi. Kah matkajad. Nemad alustasid Ventspilsist ja lõpetavad Pavilostas. Ütleks, et nõrgad, kui nad ei matkaks vastutuult. 🤣 *Puhkame ja sööme.*

K näiteks joob korraga ära üle liitri piima. Pluss muud hõrgutised. Randa tagasi jõudes hakkavad jalad juba pisut väsimusest lohisema. Aga rand on oivaline: lohesurfarid, jõesuu, päike ja lained. Koperdame mõne kilomeetri edasi ja hõivame taas ühe luitetaguse. Seekord on seal kõik kohad hõivatud.

Sipelgate poolt. Abi ei saa kutsuda, sest liinid on maas.

K surub telgi kõige madalamasse sipelgakihti, sulgeb end telki ja keeldub välja tulemast. Enne und jõuab ta märgata, kuidas tillukesed tegelased telgipinna katavad ja paar pioneeri end katusevõrgust läbi suruvad.

(Ühendkorpus 2019 osaline.blogspot.com, doc id = 3878570)

Algne mudel on andnud formaalsuse skooriks 0,38 ja spontaansuse skooriks 0,7 ning arendatud mudel annab formaalsuse skooriks 0,34 ja spontaansuse skooriks 0,61. Arendatud mudeli puhul on mõlemad skoorid vähenenud, mis on formaalsuse puhul hea märk, sest tegu päevikulaadse ja jutustava tekstiga, mis näitab, et tegu ei ole formaalsust sisaldava tekstiga. Formaalsuse hinnang on aga langenud vähe ning ise hindaksin tekstis dimensiooni veelgi madalamalt. Spontaansuse vähenemist ei saa aga heaks pidada, kuna

tekst on tüüpiline reisipäevikust blogipostitus ja sisaldab seega palju spontaansust. Kasutades aga võrdluseks teisi blogipostitustest tekste 9 ja 10, on aru saada, et tegu on neist vähem spontaanse tekstiga.

Tekst 9 on blogipostitusest reisipäeviku sissekanne, mis pärineb 2012. aasta maikuust. Tekst kirjeldab ühte päeva autori reisist Argentiinasse:

(9) *10. päev*

Päris kaua aega on läinud viimasest postitusest. Teen kiiresti enda jaanuarikuu reisi postitused ära. Hommikuks magusad argentiina viinamarjad ja mate. Enne lõunasööki võtsin päikest, mis ei olnud nii kõrvatav kui eelmistel päevadel. Lõunaks tegi tädi Betty argentiina traditsioonilist maisisuppi ehk locrot. Valmistas ühe suure potitäie värkest maisist ja lehmalihast. Juurde sai lisada nii palju juustu kui tahtsid (ikka suurtes kogustes) ja sidrunit :P Sai esimest korda seda söödud ja oli ülimaitsev:) Lõunaks shoppama Felabellasse. See on suuur shopingukeskus Cordoba kesklinnas. Oli väga moderne nagu euroopaski. Vaatasime igasuguseid riided vennale ja lõpuks osteti talle 1500 peeso eest riideid (260 eurtsi eest). Kohtasin austraalia vahetusõpilast. Täiega lahe oli kohata. Mõelda vaid, et mingis ühes poes lihtsalt kohtume. Öösel läksime kesklinna. Lollitasime ja siis saime politsei käest pahandada, kuna tahtsime monumendile ronida. jajajaj Palusin vabandust :)

(Ühendkorpus 2019 gretetangomaal.blogspot.com, doc id = 3682539)

Algne mudel on andnud formaalsuse skooriks 0,45 ja spontaansuse skooriks 0,67 ning arendatud mudel annab formaalsuse skooriks 0,43 ja spontaansuse skooriks 0,55. Ka siin on formaalsuse skoori vähenemine hea märk, sest teksti jutustav ja päevikulaadne stiil näitab, et tegu ei ole formaalse tekstiga. Hinnangu vähenemine on aga väga väike ning isiklikult hindaksin dimensiooni väärtust palju madalamalt. Seevastu spontaansuse dimensiooni väärtuse vähenemine ei ole hea. Algne mudel andnud tekstile mõõdukalt positiivse spontaansuse skoori, kuid arendatud mudel on andnud sellest palju väiksema skoori. See muutus ei ole õige, sest tegu on väga tugevalt spontaansust väljendava

tekstiga, sisaldades näiteks rohkelt emotikone, kõnekeelset sõnavara nagu *shoppama* ja ka kirjavigu nagu *kõrvatav* ja *värkset*.

Tekst 10 on 2008. aasta novembrist pärit blogipostitus Londonis elava eestlase elust:

> (10) *Kolme n2dala p2rast olengi juba varsti Eestis.. Mul on sellest vihmast nii k6rini ja tahan lund n2ha. Pole see aasta n2inudki, ainult piltide pealt. Tahan kelgutama minna, v6ib-olla suusatma ka, kuigi suurt suusatajat must ei ole, hakkavad jalad valutama :(.*
>
> *T2na siis j2lle laup2ev. T88l on toimunud m6ningased muutused. Nimelt septembrist saati on M. uues koolis. Riiklikus suures koolis, kus on igasuguseid 6pilasi ja igasuguseid 6petajaid. Paari viimase kuu jooksul on M. k2itumine totaalselt halvaks ja jubedaks muutunud ning kuna ta v2idab, et ta haugub vastu sellep2rast, et teised lapsed koolis v2idavad, et nemad teevad nii ja see on naljakas, siis jaanuarist l2heb M. privaat tydrukute kooli. Aastamakse koolis on 27 000 naela. Kujutate ette? S. sutsu nutab taga, et kui palju autosid ja kellasid ta selle raha eest osta saaks. Uues koolis on vastu haukumine vastu v6tmatu ning M. j2tab endast t2is idioodi mulje, kui midagigi teha yritab.*
>
> *Ma olen sunnitud iga v2iksemagi pisi asja yles kirjutama, kui ta halvasti k2itub. C. on hetkel nii tige, et 99%se t6en2osusega meie lapsed ei saa j6uludeks yhtegi kinki. Kujuta ette, et sa oled 6 aastane ning j6uluvana ei too sulle mitte midagi. Ma oleksin nii l88dud saades sellise r2nga 6ppetunni. Eks n2is, j6ulud ei ole enam kaugul.*
>
> *T2na 6htul l2hme Danielaga paariks tunniks minu juurde lapsi hoidma ning pyhap2eval j6ulu shopingule. Ja varsti varsti olen juba kodus.*
>
> *Yleskutse neile, kes avastavad, et neil oleks Londonist midagi vaja, siis nyyd on paras hetk m2rku anda!*
>
> *Praegu ka veel ei ole, kuid mõtlesin, et vabal hetkel võib paar sõna ju kirja panna.*

Üldiselt läheb ikka hästi. Töö on nagu on ning kurtma täna ei hakka. Pealegi olen viimastel nädalatel ekstra raha palju saanud :) (see ei tähenda, et midagi alles oleks). Ikka vanaviisi, et kolm päeva siin ning neli päeva seal.

Autol läks ükepäev soendus katki ning loomulikult olid järgmisel päeval klaasid jääs :). Nüüdseks on asi korras õnneks.

Paar nädalat tagasi käisime Thorpe Park's. Täpsemalt 31. oktoobril. Kuna samal päeval oli Halloween, oli park inimesi täis. Mega külm oli ka. 6 tunni jooksul jõudsime käia viiel atraktsioonil, igaüks neist kestis umbes 2 minutit. Ülejäänud 5 tundi ja 50 minutit seisime erinevates järjekordades. Külmetades. Üks foto ka rollercoasterilt. Kvaliteet niru, aga pole hullu.

Järgmisel päeval oli Halloweeni pidu Daniela juures. Ma ei oskagi öelda, kuidas see läks. Rohekem nalja sai dekoratsioone üles pandes ning snäkke valmistades, kui peol olles. Paljud inimesed ei ilmunud kohale ja eks me olime suht löödud ning need, kes ilmusid, sisustasid oma aega liigse alkoholi, suitsu ning narkootikumidega. Kella üheteistkümne paiku kolisin oma asjadega üles tuppa ära ning panin mingi DVD mängima. Üks pilt ka :)

Ehk siis 18-29 dets olen Eestis. Jõulude ajal ikka oma perega, ühe päeva broneerin Tikrile ja Maasikale, ühe õhtu Elinale, muu aeg on veel vaba, kuigi ma olen kindel, et mu õdedel pole ka midagi selle vastu, kui ülejäänud päevad ka nende juures olen :).

*Keegi võiks mulle lennujaama vastu ka tulla *hint hint*, hehe.*

(Ühendkorpus 2019 marimurakas.blogspot.com, doc id = 4018762)

Algne mudel on andnud formaalsuse skooriks 0,19 ja spontaansuse skooriks 0,81 ning arendatud mudel annab formaalsuse skooriks 0,16 ja spontaansuse skooriks 0,66. Nagu tekstid 8 ja 9, on ka selle blogipostituse puhul formaalsuse dimensiooni skoor parem ja spontaansus palju halvem. Formaalsuse hinnangu vähenemine on väike, kuid ka hinnang ise on väga väike, mida võib tekstist oodata tingituna märgiasendusi täis olevast esimesest

poolest. Need märgiasendused koos emotikonide ja üldiselt kõnekeelse sõnavaraga viitavad ka väga kõrgele spontaansusele, mida algne mudel on õigesti suure skooriga hinnanud. Arendatud mudeli alusel on aga tekst vaid mõõdukalt spontaansust sisaldav, mis on arusaadavalt ebasobiv hinnang.

5.3. Diskussioon

Tekstide 8, 9 ja 10 alusel saab öelda, et arendatud mudeli muutused on põhjustanud blogipostituste, kuid tõenäoliselt ka teiste rohkelt spontaansust sisaldavate tekstiliikide, liigselt väikesed spontaansuse skoorid, eriti võrreldes algse mudeliga. Kuigi nendel tekstidel on formaalsuse hinnang sobilik ning ülejäänud tekstidel on nii formaalsuse kui spontaansuse hinnangud sobilikud, on probleemne, et arendatud mudel ei anna osadele tekstidele adekvaatseid hinnanguid.

Spontaansete tekstide ebasobival hindamisel võib olla mitmeid põhjuseid. Üks põhjus võib olla kaalude arvutamiseks kasutatud hinnangukorpus, mis sisaldab vaid 180 teksti, kuid veebitekstide varieerumise uurimiseks ja juhumetsade meetodi rakendamiseks oleks aga vaja rohkem andmeid. Lisaks on selles korpuses 59 tekstil (32,8% kõikidest tekstidest) spontaansuse hinnang kõrgem kui 0,5 ja 19 tekstil (10,6% tekstidest) kõrgem kui 0,9. See tähendab, et väga spontaansuseid tekste on vähe ning seega ei ole spontaansete tekstide varieeruvus laialt esindatud. Selle probleemi lahendamiseks oleks aga vaja lisada rohkem inimhinnangutega tekste, ideaalselt mitme inimese hinnanguga, et hinnangute keskmine vähendaks üksiku hindaja subjektiivsust. See on aga väga suur lisatöö, mis nõuab eraldiseisva uurimuse tegemist.

Teine põhjus, miks kõrge spontaansusega tekstide hindamine on arendatud mudelis halvenenud, on tunnuste eemaldamine. Algses mudelis kasutatud spontaansuse tunnused – kirjavigade protsent ja EstNLTK oletamiseta morfoloogilisel analüüsil tundmatuks jäänud lühikeste sõnade protsent – on mõlemad arendatud mudelis asendatud leksikonidega. Algses mudelis olid tunnused osaliselt kattuvad, kus mõlemad tunnused põhinesid sisemistel sõnastikel ning sõnavormid võisid kirjavea või tundmatu sõna märgendi saada ainult siis, kui seda sisemistes sõnastikes ei leidunud. Selle tulemusena said paljud sõnad valepositiivse märgendi ehk märgitud spontaansust väljendavaks, kui

need seda ei olnud, nagu uudissõnad, terminitest laensõnad ning mitmed nimed ja lühendid. Arendatud mudel toimib aga vastupidiselt: valepositiivsete märgendite vältimiseks otsitakse spontaansust väljendavaid sõnu leksikonidest. Kuid leksikonid ei ole täiuslikud ning nendest puuduvad mitmesugused sõnad ja sõnavormid, eriti kirjavead, mida on ka näha tekst 7 puhul. Iga võimaliku kiiresti kirjutamisest tekkinud näpuvea leksikoni panemine ei ole võimalik, mis on suures osas tingitud sellise leksikoni mahust. On võimalik, et nii leksikonide kui ka automaatse kirjaveatuvastaja ja lühikeste tundmatute sõnade koos kasutamine annab paremaid tulemusi, kuid on teada, et need tunnused on väga kattuvad, mistõttu see ei ole lihtne.

Kolmas põhjus võib olla, et spontaansuse tunnuste loend ei ole lõplik ning tunnuste lisamine lahendaks probleemi. Kui formaalsust on varasemalt automaatselt tunnuste alusel hinnatud, siis spontaansust kui suulisele keelele sarnast kirjakeelt ei ole kasutatud automaatse hindamise objektina (vt alampeatükki 1.1). Seega tunnuseid otsides sai formaalsuse tunnustest täpsema loetelu kui spontaansuse tunnustest, ning on võimalik, et mudelist on jäänud mõni väga oluline spontaansuse tunnus välja. Kaalude arvutamisel selgus, et käändsõnade protsent on väga oluline tunnus spontaansuse hindamisel, mida ma ei olnud varasemate uurimuste põhjal bakalaureusetöös (Gailit 2021) tunnuseks määranud. Seega tasuks pärast tekstidest rohkemate tunnuste välja võtmist edasiste dimensioonide hindamiseks vaadelda kaalude arvutamise protsessis, kas tunnused on olulised spontaansuse või ka formaalsuse hindamisel.

Neljas põhjus, miks arendatud mudel annab algsest mudelist kehvemaid hinnanguid kõrge spontaansusega tekstidele, on mudelite hinnangusüsteemide erinevused. Algse mudeli hinnang on vahemikus -1 kuni 1, kus null on neutraalne tekst. Teisendatuna vahemikku null kuni üks, on neutraalseks punktiks 0,5 ning sellest kõrgema skooriga tekstid on pigem spontaansed ja madalamad tekstid ebaspontaansed. Arendatud mudel hindab tekste aga teistmoodi: 0 väljendab mittespontaansust ja 1 spontaansust. Arendatud mudel ei käsitle ebaspontaansust, mis tähendab, et mudelite hinnangud ei ole ideaalselt võrreldavad.

Ühendkorpuse 100 000 teksti suurusel valimil arvutatud tekstide keskmine spontaansus on arendatud mudeli puhul 0,36 ja algse mudeli puhul 0,60, teisendatuna vahemikku null kuni üks. Eeldades, et valimi keskmine spontaansus kattub selle keskmise teksti spontaansusega, on näha, et arendatud mudeli puhul on keskmise teksti spontaansus madalam kui algses mudelis. Sellest võib järeldada, et kõrge spontaansuse hinnang algab arendatud mudeli puhul madalamalt kui algses mudelis. Vaadates analüüsitud tekste, on näha, et blogipostitused (tekstid 8, 9 ja 10) on saanud vaadeldud tekstidest kõige suuremad spontaansuse hinnangud, mis vastab tekstide sisule. Kuid on aru saada, et need tekstid on väga tugevalt spontaansed ning väide, et tekst skooriga 0,6 skaalal null kuni üks on väga tugevalt spontaanne ei ole intuitiivselt arusaadav. Seega oleks kasulik uurida, kuidas oleks parim viis nihutada spontaansuse hinnanguid rohkem skaala otstesse, et hõlbustada hinnangutest arusaamist.

Tasuks aga täheldada, et vaadeldud tekstides on madala spontaansusega tekstide hindamine paranenud ning arendatud mudeli hinnangukorpusel arvutatud keskmine veamäär on langenud. Nende kahe aspekti tõttu saab öelda, et arendatud mudeli puhul on spontaansuse hindamine mitmekesise muutumisega.

Võrdlesin ka formaalsuse hinnanguid Ühendkorpuse valimil, kus algse mudeli keskmine formaalsuse hinnang on 0,49 ja arendatud mudeli oma on 0,56. See muutus on väiksem, suurenedes vaid 0,07 võrra, mistõttu on formaalsuse hinnangud mudelite vahel võrreldavamad. Tekstide hinnangute analüüsi alusel saab öelda, et formaalsuse hinnang on üleüldiselt paranenud, mida kinnitab ka arendatud mudeli hinnangukorpusel arvutatud keskmise veamäära langemine algse mudeliga võrreldes.

Kuid need muutused ei ole ideaalsed, mida on eriti näha kõrge spontaansusega tekstide ebatäpses hindamises. On aru saada, et spontaansuse hindamist tuleb edasi arendada, kas lisades tunnuseid, parandades kaale suurema andmestiku abil, muutes skoori vahemikku null kuni üks teisendamise algoritmi või täiustades leksikone. Lisaks võib osutuda, et mitme muudatuse sisse toomine on vajalik hinnangute parandamiseks. Viimane muudatus, leksikonide täiendamine, on aga iseeneslikult tarvilik, kuna uusi sõnu tekib pidevalt juurde kas laenamise, uue terminoloogia vajaduse või noortekeele pideva

muutlikkuse tõttu. Leksikonide täiendamine aitaks parendada ka formaalsuse hinnanguid, sest tunnus on ka selle dimensiooni hindamiseks kasutusel.

Üleüldiselt saab mudeli kohta öelda, et kuigi kõrgelt spontaansete tekstide hindamine on arendatud mudeli puhul algsest mudelist ebatäpsem, annab arendatud mudel sobivaid hinnanguid, mis on sageli paremad kui arendamata mudel. Sellest saab järeldada, et dimensiooni või selle puudumist väljendavate sõnade leksikonide kasutamine ning dimensiooni hindamiseks kasutatud tunnustele kaalude määramine on olnud kasulik ning neid aspekte oleks mõistlik kasutusele võtta ka tulevaste dimensionaalse tekstimudeli dimensioonide hindamise mudelitele.

Kokkuvõte

Töö eesmärk oli edasi arendada bakalaureusetöös (Gailit 2021) loodud dimensionaalse tekstimudeli (Vaik jt 2020) formaalsuse ja spontaansuse dimensioone hindavat mudelit. Lisasin mudelile kaks aspekti: spontaansust ja mitteformaalsust väljendavate sõnade ja sõnavormide leksikonid ning kummagi dimensiooni tunnuste olulisused ehk kaalud. Arendatud programm hindab sisendtekstide formaalsust ja spontaansust skaalal null kuni üks, kus null tähistab dimensiooni puudumist ehk mittedimensionaalsust ja üks tähistab dimensiooni tugevat esinemist. Mudel väljastab kaks arvu, millest üks on spontaansuse hinnang ja teine on formaalsuse hinnang.

Magistritöö kirjalikus osas olen kirjeldanud mudeli osasid: dimensioonide hindamisel kasutatavaid tunnuseid ning nende kaale. Formaalsuse hindamisel kasutan tunnustena sõnavara mitmekesisust; keskmist lemmapikkust; emotikonide arvu; käändsõnade protsenti; esimese, teise ja kolmanda isiku asesõnade ja verbide protsente; umbisikulise tegumoe protsenti; *nud*-partitsiibi vormis verbide protsenti; kaudse kõneviisi protsenti ning leksikonides esinevate sõnade protsenti. Spontaansuse hindamisel kasutan tunnustena sõnavara mitmekesisust, keskmist lemmapikkust, emotikonide arvu, käändsõnade protsenti, esimese isiku asesõnade protsenti, puuduva suure algustähega sõnade protsenti, läbinisti suurtäheliste sõnade protsenti, sõnasiseste korduste arvu ning leksikonides esinevate sõnade protsenti.

Kirjeldasin pikemalt kahte lisatud aspekti, leksikone ja tunnuste kaale. Spontaansust ja mitteformaalsust väljendavad leksikonid on keele- ja kirjavead, laensõnad, toorlaenud, lühendid, märgiasendused, omasõnad, partiklid, kokkukirjutamised ja Ühendsõnastikus kõnekeelsuse märgendiga sõnad. Tunnustele kaalude lisamiseks kasutasin 180 teksti suurust inimhinnangutega korpust, et treenida kummagi dimensiooni jaoks juhumetsad, millest võtsin kaaludeks %IncMSE. Formaalsuse hindamisel kasutasin saadud kaale otse, kuid spontaansuse puhul osutus kasulikuks nende juurimine 1,4-ga, et kaale ühele lähemale viia.

Viimaseks võrdlesin sisse viidud arendustega mudelit algse bakalaureusetöö mudeliga. Arvutasin mõlema jaoks keskmised veamäärad. Arendatud formaalsuse mudel sai

keskmiseks veamääraks 0,19, mis on 0,11 võrra ehk 37,06% parem arendamata mudelist. Arendatud spontaansuse mudel sai keskmiseks veamääraks 0,16, mis on 0,21 võrra ehk 55,88% parem arendamata mudelist. Need tulemused on aga arvutatud samal andmestikul, mille alusel olid kaalud määratud, mistõttu vaatlesin ka kümmet teksti, mis selles andmestikus ei sisaldunud. Vaatlesin tekstide hinnanguid käsitsi, võrreldes iga teksti puhul arendamata ehk algse mudeli ja arendatud mudeli hinnanguid.

Lisatekstide alusel osutus formaalsuse hindamine arendatul mudelil üleüldiselt paremaks ning ka madala spontaansusega tekstide hindamine paranes. Kõrge spontaansusega tekstide hindamise puhul on aga algse mudeli skoorid kõrgemad kui arendatud mudeli puhul, ehk arendatud mudel hindab kõrget spontaansust kehvemini. Selle põhjused võivad olla ebapiisav tunnuste hulk, valed kaalud või asjaolu, et arendatud mudeli Ühendkorpuse 100 000 teksti keskmine spontaansuse hinnang on palju madalam kui algse mudeli oma, ehk kõrge spontaansuse piirmäär algab palju madalamalt, kui oleks intuitiivne.

Kuigi sisse toodud muutused annavad spontaansuse hindamise puhul vastandlikke tulemusi, on formaalsuse hinnangute paranemise alusel võimalik öelda, et leksikonide kui tunnuse ning kõikidele tunnustele kaalude lisamine on dimensioonide automaatsel hindamisel kasulik ning seega võib neid sobitatud kujul kasutada ka teiste dimensionaalse tekstimudeli dimensioonide hindamisel.

Kasutatud allikad

Gailit, Karl Gustav 2021. Spontaansuse ja formaalsuse kui dimensionaalse tekstimudeli dimensioonide automaatne hindamine veebitekstides. Tartu: Tartu Ülikool.

Hennoste, Tiit 2002. Suulise kõne uurimine ja sõnaliigi probleemid. Toim. Renate Pajusalu, Ilona Tragel, Tiit Hennoste, Haldur Õim. Teoreetiline keeleteadus Eestis. Tartu Ülikooli üldkeeleteaduse õppetooli toimetised, 4. Tartu: Tartu Ülikooli Kirjastus, 56–73.

Heylighen, Francis ja Jean-Marc Dewaele 1999. Formality of Language: definition, measurement and behavioral determinants. Leo Aposteli keskuse sisearuanne. Brüssel: Vrije Universiteit Brussel.

Kasik, Reet 2007. Sissejuhatus tekstiõpetusse. Tartu: Tartu Ülikooli kirjastus.

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. – Language Resources and Evaluation 55, 757–788. https://doi.org/10.1007/s10579-020-09519-z

Laippala, Veronika, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, Sampo Pyysalo 2022. Register identification from the unrestricted open Web using the Corpus of Online Registers of English. – Language Resources and Evaluation. https://doi.org/10.1007/s10579-022-09624-1

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

Mosquera, Alejandro ja Paloma Moreda 2011. The Use of Metrics for Measuring Informality Levels in Web 2.0 Texts. – Proceedings of the 8th Brazilian Symposium in Information and Human Language Technology.

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

Muischnek, Kadri, 2015. Keelekorpused – sama mitmekesised kui keel ise. – Oma Keel 1, 37−44.

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

Rezapour Asheghi, Noushin, Katja Markert, Serge Sharoff 2014. Semi-supervised graph-based genre classification for web pages. – Proceedings of TextGraphs-9: The workshop on graph-based methods for natural language processing. Doha: Association for Computational Linguistics, 39–47.

Santini, Marina, Alexander Mehler, Serge Sharoff 2010. Riding the Rough Waves of Genre on the Web. – Genres on the Web: Computational Models and Empirical Studies. (Text, Speech and Language Technology 42.) Toim Alexander Mehler, S. Sharoff, Marina Santini. Dordrecht: Springer Publishing Company, lk 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.

Summary. Lexicons and feature weights as an addition to improve the evaluation of the dimensions of formality and spontaneity of online texts

The goal of this thesis was to experiment with the addition of lexicons and feature weights as a means to improve the evaluation of formality and spontaneity as dimensions of the dimensional text model (Vaik et al. 2020) of online texts. The practical goal of the experiment was to create a new model that improves on the evaluations made by the previous model (Gailit 2021).

To improve the model, I have added two aspects. The first is a new feature of lexicons containing words that express spontaneity and informality, such as slang, writing errors, and loanwords. The second is feature weights, where, for evaluating both dimensions, I have assigned to all of the used features a weight based on the %IncMSE of the feature from a random forest trained on a small dataset with human evaluations to predict the value of the dimension.

In comparison to the initial model, the model with the additions has improved. Its MSE has been reduced to 0,19 for formality and to 0,16 for spontaneity. In comparison to the initial model, that is 37,06% and 55,88% respectively. When analyzing the evaluations of ten previously unseen texts by the model with the additions against the initial model, the evaluation of formality has improved, as has the evaluation of low spontaneity. The evaluation of high spontaneity has, however, worsened, which could be due to inaccurate weights, an incomplete set of features or having scores considered high start at an unintuitively low minimum value.

Despite the issues with evaluating the spontaneity of texts with high spontaneity, the additions have proven to be generally useful in improving the evaluation of formality and spontaneity and thus it will be useful to include these features when developing the models to evaluate the remaining dimensions of the dimensional text model.

Lisad

Tabel 3. **Sõnavara mitmekesisuse mõju dimensioonide skooridele.**

Sõnavara mitmekesisus	Protsendid Ühendkorpuse valimis	Formaalsuse punktid	Spontaansuse punktid
0 – 0,45	0% – 9,1%	-5	+5
0,45 – 0,52	9,1% – 18,2%	-4	+4
0,52 – 0,57	18,2% – 27,3%	-3	+3
0,57 – 0,61	27,3% – 36,4%	-2	+2
0,61 – 0,65	36,4% – 45,5%	-1	+1
0,65 – 0,68	45,5% – 54,5%	0	0
0,68 – 0,71	54,5% – 63,6%	+1	-1
0,71 – 0,74	63,6% – 72,7%	+2	-2
0,74 – 0,77	72,7% – 81,8%	+3	-3
0,77 – 0,81	81,8% – 90,9%	+4	-4
0,81 – 1	90,9% – 100%	+5	-5

Tabel 4. **Keskmise lemmapikkuse mõju dimensioonide skooridele.**

Keskmine lemmapikkus	Protsendid Ühendkorpuse valimis	Formaalsuse punktid	Spontaansuse punktid
0 – 4,4	0% – 9,1%	-5	+5
4,4 – 4,54	9,1% – 18,2%	-4	+4
4,54 – 4,66	18,2% – 27,3%	-3	+3
4,66 – 4,76	27,3% – 36,4%	-2	+2
4,76 – 4,85	36,4% – 45,5%	-1	+1
4,85 – 4,95	45,5% – 54,5%	0	0
4,95 – 5,05	54,5% – 63,6%	+1	-1
5,05 – 5,15	63,6% – 72,7%	+2	-2
5,15 – 5,28	72,7% – 81,8%	+3	-3
5,28 – 5,47	81,8% – 90,9%	+4	-4
5,47 – ∞	90,9% – 100%	+5	-5

Tabel 5. **Käändsõnade protsendi mõju dimensioonide skooridele.**

Käändsõnade osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid	Spontaansuse punktid
0 – 0,33	0% – 9,1%	-5	+5
0,33 – 0,38	9,1% – 18,2%	-4	+4
0,38 – 0,42	18,2% – 27,3%	-3	+3
0,42 – 0,45	27,3% – 36,4%	-2	+2
0,45 – 0,48	36,4% – 45,5%	-1	+1
0,48 – 0,51	45,5% – 54,5%	0	0
0,51 – 0,53	54,5% – 63,6%	+1	-1
0,53 – 0,55	63,6% – 72,7%	+2	-2
0,55 – 0,58	72,7% – 81,8%	+3	-3
0,58 – 0,62	81,8% – 90,9%	+4	-4
0,62 – 1	90,9% – 100%	+5	-5

Tabel 6. **Asesõnade esimese isiku protsendi mõju dimensioonide skooridele.**

Asesõnade esimese isiku osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid	Spontaansuse punktid
0	0% – 16,7%	0	0
0 – 0,24	16,7% – 33,3%	-1	+1
0,24 – 0,5	33,3% – 50%	-2	+2
0,5 – 0,67	50% – 66,7%	-3	+3
0,67 – 0,89	66,7% – 83,3%	-4	+4
0,89 – 1	83,3% – 100%	-5	+5

Tabel 7. **Asesõnade teise isiku protsendi mõju formaalsuse skoorile.**

Asesõnade teise isiku osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid
0	0% – 50%	0
0 – 0,2	50% – 66,7%	-3
0,2 – 0,5	66,7% – 83,3%	-4
0,5 – 1	83,3% – 100%	-5

Tabel 8. **Asesõnade kolmanda isiku protsendi mõju formaalsuse skoorile.**

Asesõnade kolmanda isiku osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid
0	0% – 33,3%	0
0 – 0,2	33,3% – 50%	+2
0,2 – 0,42	50% – 66,7%	+3
0,42 – 0,8	66,7% – 83,3%	+4
0,8 – 1	83,3% – 100%	+5

Tabel 9. **Verbide esimese isiku protsendi mõju formaalsuse skoorile.**

Verbide esimese isiku osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid
0	0% – 33,3%	0
0 – 0,11	33,3% – 50%	-2
0,11 – 0,22	50% – 66,7%	-3
0,22 – 0,38	66,7% – 83,3%	-4
0,38 – 1	83,3% – 100%	-5

Tabel 10. **Verbide teise isiku protsendi mõju formaalsuse skoorile.**

Verbide teise isiku osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid
0	0% – 16,7%	0
0 – 0,06	16,7% – 33,3%	-1
0,06 – 0,13	33,3% – 50%	-2
0,13 – 0,2	50% – 66,7%	-3
0,2 – 0,33	66,7% – 83,3%	-4
0,33 – 1	83,3% – 100%	-5

Tabel 11. **Verbide kolmanda isiku protsendi mõju formaalsuse skoorile.**

Verbide kolmanda isiku osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid
0 – 0,4	0% – 16,7%	0
0,4 – 0,54	16,7% – 33,3%	+1
0,54 – 0,67	33,3% – 50%	+2
0,67 – 0,8	50% – 66,7%	+3
0,8 – 0,95	66,7% – 83,3%	+4
0,95 – 1	83,3% – 100%	+5

Tabel 12. **Umbisikulise tegumoega verbide protsendi mõju formaalsuse skoorile.**

Umbisikulise tegumoe osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid
0 – 0,02	0% – 16,7%	0
0,02 – 0,04	16,7% – 33,3%	+1
0,04 – 0,07	33,3% – 50%	+2
0,07 – 0,1	50% – 66,7%	+3
0,1 – 0,15	66,7% – 83,3%	+4
0,15 – 1	83,3% – 100%	+5

Tabel 13. ***nud*-partistiibiga verbide protsendi mõju formaalsuse skoorile.**

nud-partistiibiga verbide osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid
0	0% – 16,7%	0
0 – 0,01	16,7% – 33,3%	-1
0,01– 0,03	33,3% – 50%	-2
0,03 – 0,05	50% – 66,7%	-3
0,05 – 0,07	66,7% – 83,3%	-4
0,07 – 1	83,3% – 100%	-5

Tabel 14. **Läbinisti suurtäheliste sõnade protsendi mõju spontaansuse skoorile.**

Läbinisti suurtäheliste sõnade osakaal	Protsendid Ühendkorpuse valimis	Spontaansuse punktid
0	0% – 50%	0
0 – 0,002	50% – 66,7%	+3
0,002 – 0,009	66,7% – 83,3%	+4
0,009v– 1	83,3% – 100%	+5

Tabel 15. **Leksikonides esinevate sõnade protsendi mõju dimensioonide skooridele.**

Leksikonides esinevate sõnade osakaal	Protsendid Ühendkorpuse valimis	Formaalsuse punktid	Spontaansuse punktid
0 – 0,00008	0% – 16,7%	0	0
0,00008 – 0,002	16,7% – 33,3%	-1	+1
0,002 – 0,003	33,3% – 50%	-2	+2
0,003 – 0,005	50% – 66,7%	-3	+3
0,005 – 0,007	66,7% – 83,3%	-4	+4
0,007 – 1	83,3% – 100%	-5	+5

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Karl Gustav Gailit,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „LEKSIKONIDE JA KAALUDE LISAMINE VEEBITEKSTIDE FORMAALSUSE JA SPONTAANSUSE DIMENSIOONIDE HINDAMISE MUDELI ARENDAMISEKS“,

mille juhendajad on Kristiina Vaik ja Kadri Muischnek,

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 4.0, 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
16.06.2023