

TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Marielin Pomberg-Kalm

INFOMÕJUTUSTEKSTIDE KORPUSE KOOSTAMINE
Bakalaureusetöö

Juhendaja Kadri Muischnek, PhD

TARTU 2026

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.

Marielin Pomberg-Kalm

Lühikokkuvõte

Bakalaureusetöö eesmärgiks on informatsioonilises mõjutustegevuses kasutatavate tekstide keelekorpuse koostamine. Töö tulemusena valminud korpus koosneb olemasoleva korpuse allkorpusest ja eraldi loodud veebikorpusest. Korpuses on ligikaudu 20 000 teksti ning x sõna. Lisaks korpusele valmis töö tulemusena ka mitu üldistusvõimelist korpuse koostamist hõlbustavat töövoogu ning kontekstitundlikku märgendajat. Korpuse põhjal loodud esmane kirjeldav statistika annab alust väita, et loodud korpusel on iseloomulikke keelelisi tunnuseid.

Võtmesõnad: arvutilingvistika, korpuslingvistika, informatsiooniline mõjutustegevus

Sisukord

1.	Sissejuhatus	6
2.	Arvutilingvistika panus infokorratuse uurimisse	8
2.1.	Infokorratuse	8
2.2.	Infomõjutustekstide korpuse koostamise lähtealused	9
2.2.1.	Esinduslikkus	10
2.2.2.	Korpuse arhitektuur ja andmemudel	10
2.2.3.	Tekstide kogumine	11
2.2.4.	Tekstide järeltöötlus	11
2.3.	Infomõjutustekstide korpuse edasise arvutilingvistilise analüüsi võimalused	13
2.3.1.	Sarnased lähenemised	14
3.	Infomõjutustekstide korpuse koostamine	16
3.1.	Valim	17
3.2.	Olemasoleva veebitekstide korpuse allkorpuse loomine	18
3.2.1.	Valimisse sobivate tekstide tuvastamine ja hankimine	18
3.2.2.	Valimisse mittesobivate kirjade eemaldamine	19
3.2.3.	Täielikult kattuvate duplikaatide eemaldamine	20
3.2.4.	Osaliselt kattuvate duplikaatide tuvastamine ja esmane statistika	20
3.3.	Veebitekstide korpuse koostamine	21
3.3.1.	Valimisse sobivate tekstide tuvastamine ja hankimine	21
3.3.2.	HTML-failidest sisu ekstraheerimine	22
3.3.3.	Metaandmete leidmine ja lisamine	23
3.3.4.	Tekstielementide töötlus	23
3.4.	Tekstide puhastamine	24
3.4.1.	Artikli põhiosast välja jääva sisu eraldamine	24
3.4.2.	Vorming	24
3.4.3.	Mittestandardised teksti osad	25

3.5.	Järeltöötluse tulemused	26
4.	Märgendamine ja järeltöötlus	27
4.1.1.	EstNLTK märgenduskihid	27
4.1.2.	Mitmese analüüsiga sõnade ühestamine	28
4.1.3.	Nimeüksuste tuvastamine ja järeltöötlus	29
5.	Esialgne kirjeldav statistika	30
5.1.	Sõnaliikide jaotus	31
6.	Kokkuvõte	32
6.1.	Piirangute kirjeldus	32
7.	Kirjandus	33
Lisad	38	
Lisa 1.	Tekstitöötlusalgoritmid	38
Lisa 2.	Nimeüksuste märgenduskihi järelmärgendaja NERPostTagger	39

1. Sissejuhatus

Siinse rakendusliku suunitlusega bakalaureusetöö eesmärgiks on luua eeldused eestikeelsetes tekstides esineva informatsioonilise mõjutustegevuse arvutilingvistiliseks analüüsiks. Töö tulemusena valmis keelekorpus (edaspidi infomõjutustekstide korpus), mis sisaldab teadaolevalt informatsioonilises mõjutustegevuses osalenud allikates vahemikus 2014-2022 avaldatud meediatekste. Korpuses on 18 611 dokumenti ning selle suuruseks on 6 853 017 sõna. Lisaks on korpusele tehtud esmane kirjeldav analüüs.

Mõjutustegevus hõlmab erinevaid avaliku arvamuse, hoiakute, käitumise ja otsustusprotsesside mõjutamiseks kasutatavaid meetodeid (Kaitsepolitseiamet 2014: 13; Vabariigi Valitsus 2023: 10; Teabeamet 2017: 16). Selle alaliigiks on informatsiooniline mõjutustegevus (ingl *information influence activities* e IIA, *information influence operations* e IIO), ehk strateegiline kommunikatsiooniviis, mis kasutab eelmainitud eesmärkide saavutamiseks teabe levitamist või tõlgenduskontekstiga manipuleerimist (Ventsel & Madisson 2024: 226). Tegemist on mõjutustegevuse kõige sagedasema alaliigiga (Lange-Ionatamishvili & Mikulski 2024: 11).

Informatsioonilist mõjutustegevust ja selle ohtlikkust on julgeolekuasutuste avalikes väljaannetes käsitletud juba pikemat aega ning Kaitsepolitseiamet (2025: 22) peab seda kaudseks julgeolekuohuks ka viimati avaldatud aastaraamatus. Nähtusele on tähelepanu pööratud ka Eesti julgeolekupoliitikas (Vabariigi Valitsus 2023: 7, 10), mis peab oluliseks informatsioonilise mõjutustegevuse tuvastamist ja leviku piiramist. Tegemist on suunaga, millele on viimase kümnendi jooksul ka teaduses hakatud oluliselt rohkem tähelepanu pöörama. Näiteks on meediasisu uurijad leidnud, et isegi kui eelmainitu ei mõjuta inimeste uskumusi, siis mõjutab see nende mõtteainet, olles ühiskonnale potentsiaalseks ohuks (Lazer jt 2018: 8).

Informatsioonilises mõjutustegevuses on eesmärkide saavutamiseks kasutusel keelelised vahendid. Sellest tulenevalt on autor seisukohal, et arvutilingvistika võimaldab interdistsiplinaarse valdkonnana nähtust sügavuti uurida ning erinevaid lähenemisi kombineerida. Eestikeelsetes tekstides esinevaid mõjutustegevusele viitavaid ilminguid ei ole autorile teadaolevalt varem arvutilingvistiliste vahendite abil uuritud. Teema aktuaalsus

seisneb keele kui mõjutustegevuse vahendi uurimises ning arvutilingvistilise uurimismeetodi arendamises.

Töö põhieesmärgiks on infomõjutustekstide korpuse koostamine. Selle abil on võimalik kontrollida töö nullhüpoteesi, et informatsioonilise mõjutustegevuse keelekasutus erineb muul eesmärgil loodud meediatekstide omast. Töö põhieesmärgiks on koostada hüpoteesi kontrollimist võimaldav keelekorpus. Kitsamaks eesmärgiks on korpuse põhjal kirjeldava statistika loomine, mille abil saab nullhüpoteesi osas esmased järeldused teha.

Kirjeldav statistika hõlmab tekstide tähistustasandit ehk lausete ja tekstide keskmisi pikkusi ning sõnaliikide jaotust korpuses. Töö kitsamast eesmärgist lähtudes on püstitatud järgmine uurimisküsimus:

- Kas ja millisel viisil infomõjutustekstide korpus võrdluskorpustest¹ erineb?

Töö koosneb kuuest peatükist:

- **Teoreetilised lähtekohad** (peatükk 2) annavad ülevaate arvutilingvistika panusest informatsioonilise mõjutustegevuse uurimisse. Kirjeldatud on korpuse loomise aluseid, tulevase analüüsivõimalusi ning sarnaseid korpuseid.
- **Infomõjutustekstide korpuse koostamise** (peatükk 3) osa hõlmab uurimismaterjali tutvustust, andmekogumismeetodit ja töötluse kirjeldust.
- **Märgendamise ja järeltöötluse** (peatükk 4) osa hõlmab kasutatud märgenduskihtide ning nende järeltöötluse lühikirjeldust.
- **Tulemuste esituse** osa (peatükk 5) sisaldab valmis korpuse kohta loodud kirjeldavat statistikat.
- **Kokkuvõte** (peatükk 6) sisaldab piirangute kirjeldust ja edasisi töösuundi.

Kirjutamisel ei ole kasutatud generatiivsete keelemudelite abi.

¹ **Võrdluskorpus** – sama registri ja ajavahemiku tekste sisaldav korpus (Ädel 2020: 17)

2. Arvutilingvistika panus infokorratuse uurimisse

2.1. Infokorratatus

Kaheldava tõeväärtuse või läbipaistmatu eesmärgiga tekstide iseloomustamiseks kasutatakse sageli terminit valeuudis (ingl *fake news*). Samuti kasutatakse termineid võltsuudis ja libauudis, olenevalt käsitlusest sünonüümidenä või seotud tähendustes. Nii teaduses kui poliitikas määratletakse nähtust sõltuvalt allikast erinevalt (Aïmeur, Amri & Brassard 2023: 7), kuid on võimalik eristada kitsamat ja laiemat tähendust. Kitsamas tähenduses viitavad valeuudised väärteavet sisaldavatele ajakirjandustekstidele, laiemas tähenduses tähistab termin igasugust väärteavet või muul viisil kahjulikku teavet. Lisaks kasutatakse terminit ka soovimatut või häirivat teavet diskrediteeriva sildina. Erinevad terminid, määratlused ja kasutusviisid näitavad selgelt, et tegemist ei ole üheselt mõistetava nähtusega (*ibid.*).

Wardle ja Derakhshan (2017) pakuvad välja katusermini infokorratatus (ingl *information disorder*), mis tähistab ohtliku või eksitava info levikut ning hõlmab kolme tüüpi informatsiooni. Autorid on loonud ka infokorratust käsitleva kontseptuaalse raamistiku, milles eraldatakse infokorratuse kolme tüüpi, elementi ja faasi.

Infokorratuse tüüpideks on väär-, kuri- ja valeinfo (*ibid.* 20). Kontseptsioone on võimalik eristada lähtuvalt nende tõepärasusest ja edastaja kavatsusest:

- (1) väärinfo (ingl *misinformation*) on „tahtmatult ebatõene või eksitav“;
- (2) kuriinfo (ingl *mal-information*) on „tõene info, mida jagatakse kahju tegemise eesmärgil“;
- (3) valeinfo (ingl *disinformation*) on „teadlikult ebatõene või eksitav“. (Militerm)ⁱ

Siit järeldub, et kahjustav informatsioon ei pruugi tingimata olla väär, kuid seda saab esitada valikuliselt või fakte moonutades. Samuti ilmneb, et väär informatsioon ei pruugi tingimata olla kahjulik. Eelnevat silmas pidades võib eeldada, et kõik infotüübid on mõjutustekstides esindatud. Sellest tulenevalt ei annaks valeuudiste uurimisega piirdumine nähtusest piisavalt mitmekülgset ülevaadet, samuti ei saa loodavat korpust pidada valeuudiste korpuseks. Kuigi valeuudised on informatsioonilise mõjutustegevuse vahendiks, ei levitata valeuudiseid ainult informatsioonilise mõjutustegevuse eesmärgil ning informatsiooniline mõjutustegevus ei piirdu valeuudistega.ⁱⁱ

2.2. Infomõjutustekstide korpuse koostamise lähtealused

Infokorratuse ilminguid on võimalik arvutilingvistiliste vahenditega mitmekülgsest uurida. Korpuslingvistika ülesanded võib üldistades jagada kolmeks: korpuse koostamine, sellest teabe hankimine ja selle analüüs (Paquot & Gries 2020: v). Siinne töö keskendub korpuse loomisele ning selle edasisi analüüsivõimalusi on kirjeldatud alapeatükis 2.4.

Korpuse koostamine hõlmab selle disainimist, tekstide kogumist ja töötluks sobivasse formaati viimist, asjakohaste metaandmete kogumist ja hoiustamist ning tekstide märgendamist (Paquot & Gries 2020: 3). Kui eesmärgiks on tuvastada loodavale korpusele iseloomulikke keelelisi tunnuseid, kasutatakse ka ühte või mitut võrdluskorpust (Ädel 2020: 17).

Infomõjutustekstide korpuse koostamisel saab lähtuda Rubin, Chen ja Conroy (2015: 2) sisulisi, vormilisi ja rakenduslikke aspekte hõlmavatest nõuetest:

1. **sisu** koha pealt on oluline kasutada ka usaldusväärsetest allikatest pärit võrdluskorpust ning tagada, et sihtkorpusesse valitud allikate puhul oleks nende vastavus vaeuudise kriteeriumitele kindlalt tõendatud;
2. **vormilt** peab korpus olema masinloetaval kujul ehk tekstiformaadis (piltidel tekst juures, video/audio transkribeeritud) ning tekstid sarnase pikkusega;
3. **esinduslikkus**: uudiste tüübid ja teemad on võrdselt jaotatud, tekstid pärinevad sama tüüpi autoritelt, on erinevate allikate lõikes võrreldavad ning kindlast ajavahemikust;
4. **ressursikulu ja piirangud**: tuleb arvestada tekstide hankimiseks kuluva ressursiga ning võimalike piirangutega (autoriõigused, tundlikud andmed);
5. **kontekst**: analüüsis arvestatakse tekstide esitusviisi, keele ja kultuuriruumiga.

2.2.1. Esinduslikkus

Esinduslik valim tagab populatsiooni kohta tehtavate järelduste usaldusväärsuse. Esinduslikkuse tagamiseks on esmatähtis vaadeldav populatsioon selgelt defineerida ning valim läbimõeldult moodustada. (Biber 1993: 243)

Kindlat teemat või keelenähtust puudutava korpuse korral on tegemist erikorpusega (Pilvik jt 2025: 58), millest lähtuvalt rakendub esinduslikkuse nõue ainult valdkonnapõhistele tekstidele. Meediatekstide puhul tähendab see, et korpus võib koosneda ainult meediaväljaannetes avaldatud veebitekstidest. Kuna veebitekstide korral pole võimalik populatsiooni suurust täielikult tuvastada (Kehoe 2020: 336), saab piirduda levinumate väljaannetega. Samuti suurendab kindlatele veebilehtedele keskendumine tõenäosust, et korpus on esinduslik (Kehoe 2020: 340)

2.2.2. Korpuse arhitektuur ja andmemudel

Korpuse arhitektuur hõlmab sisu omavahelisi seoseid ja nende planeerimist. Makrotasandil koosneb korpus dokumentidest, millest igaüks moodustab keeleandmeid sisaldava objekti. Dokumentidel võib olla üks juurkaust, kuid need võivad olla ka kategooriate põhjal eraldi kaustadesse jaotatud. Levinuim on hierarhiline puustruktuur. (Zeldes 2020: 49, 51)

Siinse töö raames loodav korpus on samuti puustruktuuriga ning kategooriate põhjal jaotatud. Kirjed on JSON-failides ning töötluse ajal käsitletakse iga korpust ja allkorpust andmetabelitena, kus iga rida vastab ühele tekstile ning veerg ühele teksti kohta käivale parameetrile.

Korpuse arhitektuuri mudelid erinevad selle poolest, millisel määral neis algandmete originaalkuju säilib (*ibid.* 52). Kui teksti originaalkuju säilimine on analüüsi jaoks oluline, siis saab korrigeerida automaattöötluse jaoks teksti koopiat vormingut jättes originaalkuju muutmata (*ibid.* 53). Originaalkuju võib muutuda ka automatiseeritud tokeniseerimisel², millest tulenevalt võib väljund vajada järeltöötlust (*ibid.* 54). Loodavas korpuses säilib iga teksti algkuju (veerg *p*), mida saab hiljem võrrelda teksti puhastatud ja töödeldud kujuga (veerg *p_mod*).

² Tokeniseerimine ehk sõnestamine – teksti tähenduslikeks üksusteks tükeldamine (Pilvik jt 2025: 46)

2.2.3. Tekstide kogumine

Meediasisu on leitav spetsiaalsetest veebikorpustest. Veebikorpused koostatakse veebi automaatlehitsejate (ingl *web crawler*) abil, mis laevad veebitekstid HTML-kujul alla ning ekstraheerivad failist teksti koos muu asjakohase teabega (Kehoe 2020: 338).

Standardne veebilehtedest korpuse loomise töövoog koosneb mitmest etapist. Peale dokumentide kogumist keskendutakse hõlmab veebilehe selliste tüüpelementide eemaldamisele, mis ei ole osa veebilehe põhisistust. Järgmiseks jäetakse välja valimisse mittesobivad ning vajaduse korral ka keskmisest oluliselt erineva pikkusega tekstid. Lisaks eemaldatakse tekstide kogumise käigus tekkinud duplikaadid, mis võivad muul juhul tulemusi mõjutada. Seejärel on võimalik korpust tokeniseerida, lemmatiseerida ja morfoloogiliselt märgendada. (Kehoe 2020: 339–340)

Igas veebikorpuse loomise etapis tehtud otsused mõjutavad selle lõplikku kuju nii suuruse kui sisu mõttes. Töövoogu luues on vaja teha kindlaks, mis kuulub tüüpelementide alla, kui sarnased dokumendid selleks olema peavad, et neid enamjaolt kattuvateks duplikaatideks pidada ja kas duplikaatide otsimine peaks toimuma lause, lõigu või dokumendi tasandil. (Kehoe 2020: 340) Loodavas korpuses otsitakse duplikaate kõigil kolmel tasandil. Lisaks leitakse ka osaliselt kattuvad duplikaadid.

Korpuse juures on väga olulised metaandmed, mis lihtsustavad keelekasutust mõjutavate tegurite analüüsi ning võimaldavad iga korpuses sisalduva teksti kohta lisateavet talletada (Ädel 2020: 11, 14, 21).

2.2.4. Tekstide järeltöötlus

Korpuse esinduslikkuse tagamiseks oli vaja hinnata eelmainitud tegevuste tulemusena saadud korpuste kvaliteeti. Kvaliteedinõuetele vastava korpuse korral saab jätkata tulemuste kirjeldamise ja andmeanalüüsiga, kuid vastasel juhul on vaja pöörata tähelepanu korpuse kvaliteedi tõstmisele. Kvaliteedile rõhu asetamine on eriti oluline arvestades asjaolu, et mitteesinduslik korpus moonutab selle alusel keelest kujunevat ettekujutust (Pilvik jt 2025:

12). Sellest järeldub, et tegemist on vältimatu vaheetapiga juhul, kui kvaliteet jääb alla sihttaseme.

2.3. Infomõjutustekstide korpuse edasise arvutilingvistilise analüüsi võimalused

Infokorratuse loomiseks kasutatavate tekstide tuvastamine on paradigmaatiliselt teabe-, konteksti- või stiilipõhine. Teabepõhine lähenemine võrdleb infootsingu meetodite või teadmusbaaside abil uuritava teabe kooskõllalisust olemasoleva usaldusväärse teabega. Kontekstipõhine lähenemine kasutab sotsiaalse võrgustiku analüüsi meetodeid ning hindab teabe usaldusväärsust metaandmete ja levikumustrite põhjal. Stiilipõhine lähenemine hõlmab lause tasandil tõepärasuse hindamist (ingl *deception detection*) ning selle alaliigina hoiakute tuvastamist. (Potthast jt 2018: 232)

Kuigi Potthast jt (2018: 233) liigitavad ka tekstide kategoriseerimise stiilipõhise lähenemise alla, peetakse seda arvutilingvistikas konsensuslikult laiemaks ülesandeks. Keeleteadusliku lähenemise korral analüüsitakse tekste eesmärgiga leida neile iseloomulikke keelelisi mustreid (Conroy, Rubin & Chen 2015: 6). Samuti on võimalik kasutada erinevaid lähenemisi kombineerivaid hübriidlahendusi, mis peaksid vastama järgmistele tingimustele (*ibid.* 9):

- keeleline töötlus peaks hõlmama teksti erinevaid tasemeid;
- keeleteadusliku lähenemise korral tuleks kasutada sotsiaalsõrgustiku analüüsi usaldusväärsete allikate tuvastamiseks ja seeläbi töö usaldusväärsuse tõstmiseks;
- kasutatavad meetodid peaksid inimlikku kaalutlusvõimet toetama, mitte seda asendama ning väljundandmete ja meetodite vahelised seosed olema läbipaistvad;
- tulemused peaksid olema kuldstandardina³ avalikult kättesaadavad.

NATO Strateegilise kommunikatsiooni kompetentsikeskuse ja Euroopa hübriidohtude vastase tippkeskuse loodud raamistikus infomõjutusoperatsioonide tuvastamiseks tuuakse võimaliku uurimissuunana välja kommunikatsioonitehnikate ja konteksti analüüs, millest viimane hõlmab geopoliitilist konteksti, narratiive ja diskursust ning üldist keelekasutust (Pamment & Smith 2022: 21, 31). Barbu jt (2024) peavad loomuliku keele töötlust üheks strateegilise kommunikatsiooni analüüsi vahendiks leides, et eesti keele jaoks tuleks arendada teemade modelleerimise (ingl *topic modelling*) ja narratiivide tuvastamise (ingl *narrative detection*) meetoodika. Infomõjutustekstide korpuse koostamine võimaldab tulevikus selleni jõuda.

³ **Kuldstandard** (ingl *gold standard / golden standard*) – Käsitsi märgendatud korpus, mida kasutatakse hiljem märgendustööriista tulemuste võrdlemiseks (Pilvik jt 2025: 45)

2.3.1. Sarnased lähenemised

Seotud kirjanduses käsitletakse nii sarnastel põhimõtetel loodud korpusi kui ka nende edasisi rakendusi. Enamasti esitatakse tulemusi koos, seega lisaks loodud korpusele tutvustatakse ka selle põhjal treenitud klassifitseerimismudelit või meetoodikat. Sarnased korpused saab sisu põhjal jaotada sageduse järjekorras valeuudiste korpusteks, vaenukõne korpusteks ja konkreetselt informatsioonilisele mõjutustegevusele keskenduvateks korpusteks.

Enamus sarnaseid korpusi keskenduvad valeuudistele. Olemasolevad valeinfo tuvastamise eesmärgil loodud andmestikud sisaldavad vastavalt eesmärgile väljendeid, lauseid või terviktekste ning enamasti ka metaandmeid. Andmed pärinevad sotsiaalmeediast, poliitilistest sõnavõttudest ja uudistest. (Oshikawa, Qian & Wang 2020: 2–3)

Ühe kategooria moodustavad andmestikud, kus iga artikkel on kindlaks määratud kategooriate põhjal klassifitseeritud. Saksakeelne GerDISDETECT (Schütz jt 2024) koosneb uudistekstidest, millele on määratud vasted viies kategoorias. Hinnatakse üldist hoiakut (*General View*), teksti stiili (*Writing Style*), teabe esitamise viisi (*Reporting Style*) ning vihakõnele (*Hate Speech*), ja ekstremismile (*Extremism*) viitavate tunnuste esinemist (*ibid.* 7683). Lause tasandil on võimalik tuvastatada uudistekstides sageli inimeste mõjutamiseks kasutatavaid diskursiivseid tehnikaid (Yu jt 2021).

On loodud ka selliseid andmestikke, mis keskenduvad konkreetselt informatsioonilisele mõjutustegevusele. Eelmisel aastal loodi Balti riikides esimene infomõjutustekstide korpus HALT-PROP (Rizgeliene jt 2025), mis on leedukeelne ning sisaldab 2870 käsitsi klassifitseeritud artiklit ning 1000 infomõjutustehnikate ja narratiivide suhtes käsitsi märgendatud artiklit. Üldisema fookusega on poolakeelne MIPD (Modzelewski jt 2024), mille autorid määrasid artikli temaatilise kategooria, hindasid artikli algallika ja autori usaldusväärsust, klassifitseerisid artikli ning uurisid desinformatsiooni korral manipuleerimist, kavatsusi ja narratiive.

Infomõjutustekstide korpuste põhjal on loodud mitmeid tööriistu, millest enamus keskenduvad Kremli mõjutustegevuse tuvastamisele. Tavakasutajale on kättesaadav näiteks Check News in 1 Click (Solopova jt 2024), mille loomiseks kasutati faktikontrolli läbinud artiklid, 5% sisust

oli teadaolevalt Kremliga seotud poliitiliste parteide pressiteated ning uudise usaldusväärsuse hindamisel kasutati morfosüntaktilisi tunnuseid ning võtmesõnu (*ibid.* 45).

Eraldi kategooriana saab välja tuua ka Ukraina sõjaga seonduvad andmestikud. Sellistest andmestikest on loodud näiteks ingliskeelne korpus WarMM-2022 (Alyukov, Kunilovskaya & Semenov 2023), millega uuritakse võtmesõnapõhise korpusanalüüsi abil Ukraina sõjaga seonduvat informatsioonilist mõjutustegevust sotsiaalmeedias. Samuti on loodud mitmekeelne uudistekstide ja Telegrami-sõnumite korpus (Solopova jt 2023), mille põhjal arendati mitu automatiseeritud meetodit kremlimeelse sisu tuvastamiseks.

3. Infomõjutustekstide korpuse koostamine

Osas 3.1 kirjeldan varem määratletud kriteeriumite põhjal koostatud valimit. Osades 3.2 ja 3.3 kirjeldan allkorpuste loomise töövooge, mis toetuvad eespool esitatud teoreetilistele lähtekohtadele. Seejärel kirjeldan osas 3.4 loodud allkorpuste puhastamiseks loodud töövoogu. Viimaks kirjeldan osas 3.5 loodud allkorpuste ühendamist.

Minul oli läbi Tartu Ülikooli ligipääs eestikeelsete veebitekstide korpusele (edaspidi **Ko1**) ning kõrgjõudlusega keeletehnoloogia (High Performance Language Technologies, HPLT) projekti (Burchell jt 2025) raames avaldatud Common Crawl korpuse versioonile (edaspidi **Ko2**). Olemasolevate veebikorpuste allkorpuste loomise töövoog osas 3.2 hõlmab tegevusi korpustes sisalduvate valimisse sobivate tekstide tuvastamisest täielikult kattuvate duplikaatide eemaldamiseni.

Töö käigus ilmnas, et mõlemas allkorpuses esines osaliselt kattuvaid duplikaate. Sellest tulenevalt oli vajalik kontrollida tekstide vastavust algallikas avaldatule. Korpuse suuruse tõttu tuli protsess automatiseerida, mistõttu kujunes välja ka teine töövoog. Originaalkujul veebitekstide korpuse (edaspidi **Ko3**) loomise töövoog osas 3.3 hõlmab valimisse sobivate veebiaadresside tuvastamist, neile vastavate failide hankimist, valimisse sobivate tekstide tuvastamist, failis sisalduvate metaandmete lisamist ning failist tekstisisu ekstraheerimist.

Osas 3.4 kirjeldan eespool mainitud osaliselt kattuvate duplikaatide arvu vähendamise eesmärgil tekstide puhastamist ja vormingu korrigeerimist. Teatud juhtudel erinesid tekstid tähemärgi tasemel, mil sai tekstide vormingut korrigeerides need täielikeks duplikaatideks teisendada. Enamasti olid osalise kattuvuse põhjuseks Ko1 ja Ko2 tekstide automaatse kogumise käigus alles jäänud osad, mis ei kuulunud artikli teksti hulka. Töövoog kasutab reeglipõhist sõnetöötlust ja loomuliku keele töötluste tarkvara EstNLTK. Olemas on ka lisateekidest sõltumatu varuvariant.

Kirjeldan osas 3.5 Ko1, Ko2 ja Ko3 ühendamist ja duplikaatide eemaldamist.

3.1. Valim

Korpuse koostamisel oli alapeatükis 2.3.1. kirjeldatust lähtuvalt esimeseks sammuks valimi selge defineerimine. Vastavalt Rubin, Chen ja Conroy (2015: 2) infomõjutustekstide korpuse valimi kriteeriumitele pidi infomõjutustegevuseks kasutatavate tekstide korral nende sobivus valimisse tõendatav olema. Sellest tulenevalt lähtus allikate valik olemasolevast kirjandusest.

Allikate valik

Valim sisaldab järgnevaid kategooriaid:

K1 (IMT) – Teadaolevad mõjutustegevuses osalevad allikad

K2 (ALT) – Alternatiivmeedia

K3 (ERR) – Eesti Rahvusringhääling

Infomõjutustekstide korpus (K1-IMT), siinse töö põhikorpus, sisaldab teadaolevalt mõjutustegevuses osalevate väljaannete tekste. Tegemist on väljaannetega, mida mainitakse julgeolekuasutuste avalikes materjalides, tippkeskuste raportites ja teemakohases akadeemilises kirjanduses.

Lisakorpusena on kasutusel alternatiivmeedia tekstikorpus (K2-ALT). Selle eesmärgiks on võrrelda mõjutustekstide ja alternatiivmeedia tekstide keelekasutust. Alternatiivmeedia allikate valik lähtus Globaalse Valeformatsiooni Indeksi (GDI; Fagan, Jõesaar & Kõnno 2020) raportist Eesti kohta ning 2023. aasta Eesti ühiskonna lõimumismonitooringust (Voog jt 2024).

Võrdluskorpus (K3-ERR) sisaldab Eesti Rahvusringhäälingu eestikeelses portaalis avaldatud tekste. Valik lähtus mitmest asjaolust. Esiteks on ERR GDI (2020: 6) andmetel üks kõige madalama valeinfo leviku riskiga väljaandeid. Teiseks soodustas valikut asjaolu, et tegemist on avalik-õigusliku asutusega, millest tulenevalt ei ole artiklitele ligipääsupiiranguid.

Ajavahemik

Valim sisaldab tekste vahemikus 2014–2022 (k.a.), mis hõlmab perioodi mõjutustegevuse läänemaailma suhtes ohuna käsitlemise algusest aastal 2014 (Kaitsepolitseiamet 2014: 4) kuni Ukraina sõja alguseni, mil teadaolevate mõjutusallikate sisu avaldamine sanktsioonide tõttu katkes.

3.2. Olemasoleva veebitekstide korpuse allkorpuse loomine

Olemasoleva veebitekstide korpuse allkorpuse loomise töövoog hõlmab kasutusel olevates korpustes sisalduvate valimisse sobivate tekstide tuvastamist, nende hankimiseks korpuspäringute tegemist, tekstide töötluks sobivale kujule viimist, valimisse sobivuse kontrolli ja ebasobivate kirjete välja jätmist ning täielikult kattuvate duplikaatide eemaldamist.

3.2.1. Valimisse sobivate tekstide tuvastamine ja hankimine

Määrasin alapeatükis 3.2 kirjeldatud valimi koosseisu põhjal vaadeldavatele väljaannetele vastavad domeeninimed ning sain nende põhjal olemasolevatest korpustest vastavad tekstid.

IMT kategooria tekstidest olid korpustest kättesaadavad nelja erineva allika tekstid (allkategooriad K1_1 kuni K1_4), kategooria 2 puhul piirdusin samuti nelja kõige laiemalt levinud allikaga. Kategooria 3 puhul oli ajamärgistatud Ko1 ERR-i tekstide hulk piisavalt suur, et ei vajanud lisapäringuid Ko2-st.

K1-IMT tekste oli töötlemata kujul kokku 23 442, millest pea kõik (97%, n=23 442) olid allikast K1_1. Samas kategoorias oli veel kolm allikat, millest igaühes oli ligikaudu 200 teksti (K1_2: n=274, K1_3: n=181, K1_4: n=220).

3.2.2. Valimisse mitesobivate kirjete eemaldamine

Selles etapis võetakse toorandmed, viiakse kõik tabelid ühele kujule ning võimalusel ekstraheeritakse veebiaadressidest (ingl *uniform resource locator* e URL, edaspidi kasutusel lühend) metaandmed. Tabelite ühele kujule viimine hõlmab veergude ümbernimetamist ja lisamist ning andmetüüpide muutmist. Järgmiseks jäetakse ekstraheeritud metaandmetest või muudest määratud tingimustest lähtudes välja kõik valimi kriteeriumitele mittevastavad kirjed.

Seejärel jätan alles ainult töötamiseks vajalikud veerud. Järgmiseks hindan kirjete sobivust valimisse. Mitesobivus võib tuleneda avaldamisajast või sisu tüübist. Kõigil tekstidel oli olemas veebist automaatselt kogumise kuupäev (metaandmete parameeter *crawl date*), mis võis aga avaldamisajast mitme aasta võrra erineda. K2 ja K3 puhul oli uudisvoost kogutud tekstide hulk piisavalt suur, et ilma lisatöötluseta voo ajamärgistuse põhjal valimisse sobivad tekstid eraldada. K1-IMT puhul oli teavet ebapiisaval määral, mistõttu töötasin välja eraldi meetodi tegelike avaldamisaegade leidmiseks.

Kui URL sisaldas kuupäeva, sai selle põhjal avaldamisaja päeva täpsusega leida ning ajavahemikku sobivust kontrollida. Kui avaldamisaega polnud võimalik URL-i põhjal tuvastada, leidsin iga kirje jaoks algallikast URL-i järgi artikli, kontrollisin seal olevat avaldamisaega ning määrasin sellest lähtuvalt valimisse sobivate ID-de vahemiku. Kui avaldamisaega polnud võimalik määrata või ligipääs algallikale puudus, jäi tekst valimist välja.

Samuti sai URL-i põhjal teatud juhtudel määrata sisu tüübi. Kuna korpuse kvaliteedi nõuded näevad vastavalt alapeatüki 4.1 punktide 2 ette, et muus formaadis sisu on vaja transkribeerida ning siinse töö maht ei oleks meediasisu teksti kujule viimist võimaldanud, jäid välja ka kõik sellised tekstid, mille sisu oli põhiliselt pildi, video, helifaili või muu graafilise esituse kujul. Kui automaatse kogumise käigus oli ühe teksti alla mitu teksti pandud (92 juhtu), siis jäi alles ainult esimene.

K1_1 tekstidest jäid enamus alles (85,7%; n=20 086), K1_2 tekstidest ligikaudu pool (54,4%; n=149), K1_3 tekstidest alla poole (36,5%; n=66) ning K1_4 tekstidest kõik (n=220).

3.2.3. Täielikult kattuvate duplikaatide eemaldamine

Kuigi absoluutne URL ehk terve internetiaadress oli esialgsetes korpustes unikaalne, võisid aadressid teatud juhtudel tegelikult kattuda. See võis tuleneda erinevast pöördusprotokolli nimest (HTTP/HTTPS), domeeninimest (tava- ja mobiiliversioon) ja URL-is sisalduvast päringust. Kõigil neil juhtudel käsitles tekstide automaatseks kogumiseks kasutatud tarkvara neid eraldi aadressidena.

Võrdlesin selliste kirjete tuvastamiseks URL-ide pöördusteid (ingl *path*). Topeltkirjete tuvastamine lähtub mitmest mõõdikust ning toimub mitmel tasandil. Kirjed grupeeritakse pöördustee järgi ning seejärel kontrollitakse mitme kirje esinemise korral sisu kattuvust. Kui sisu ühtib täiel määral, siis jäetakse allkorpuse tasandil alles esimene kirje ning ühendkorpuse tasandil olemasolu korral Ko1 kirje. Kui grupeeritud kirjed täiel määral ei kattu, tehakse selle kohta vastav märge ning kattuvust kontrollitakse uuesti etapis 3.5.1 peale tekstide puhastamist.

Täielikud duplikaadid esinesid ainult Ko2 tekstide hulgas. K1_1 tekstidest jäi alles 94,6% (n=4813), K1_2 tekstidest 98,9% (n=89), K1_3 tekstidest 92,4% (n=61), K1_4 tekstidest 98,5% (n=203).

3.2.4. Osaliselt kattuvate duplikaatide tuvastamine ja esmane statistika

Osalist kattumist põhjustas pealkiri, avalause või artikli sisu. Oli ka juhte, mil mitu eelmainitust omavahel erinesid. Pealkiri võis sisaldada näiteks väljaande nime, avalause olla pealkirjaga asenduses ning tekstis jäänud sellised *p*-isendid, mis ei kuulunud artikli teksti hulka.

Allikas		Tekste kokku	Unikaalseid	Osakaal
K1_1	Ko1	14 996	14 402	96%
	Ko2	5090	4766	93,6%
K1_2	Ko1	59	56	94,9%
	Ko2	90	87	96,7%
K1_3	Ko2	66	61	92,4%
K1_4	Ko1	14	14	100%
	Ko2	206	203	98,5%

Tabel 1. Võtmeveeru *url_path* väärtuste unikaalsus

3.3. Veebitekstide korpuse koostamine

Veebitekstide korpuse loomise töövoog hõlmab valimisse sobivate väljaannete tuvastamist, nende sisu kättesaadavuse kontrolli, olemasolevate failide allalaadimist, failidest teksti ja muu asjakohase sisu ekstraheerimist ning järeltöötlust.

Protsessi automatiseerimiseks on mitmeid tööriistu, kuid antud juhul otsustasin neid mitte kasutada. Uuritavate väljaannete töötlemiseks spetsiaalse lähenemise loomine võimaldab vähendada automaattöötlustest tulenevat andmekadu ning tagada suurema esinduslikkuse määra.

Kasutasin HTML-failide parsimiseks⁴ ja töötlemiseks teeki BeautifulSoup (Richardson 2025). HTML-failid on hierarhilise puustruktuuriga ning teegi abil on võimalik struktuuri muuta.

3.3.1. Valimisse sobivate tekstide tuvastamine ja hankimine

Tegin API-päringuid, et leida kõik arhiveeritud URL-id, mis sisaldavad mõnda valimisse kuuluvatest domeeninimedest. Seejärel filtreerisin nende hulgast faililaiendi või muu asjakohase mustri põhjal välja sobiva sisutüübiga failid.

Peale sobivate URL-ide määramist tegin pakktöötluste⁵ põhimõttel API-päringuid igale aadressile vastava HTML-faili alla laadimiseks.

Tabel 2. Ülevaade sobivate URL-ide arvu muutumisest protsessi käigus.

	Allikas			
	K1_1	K1_2	K1_3	K1_4
Domeeninime sisaldavad URL-id	54 227	30 581	1 112 579	3635
Unikaalsed URL-id	35 266	19 652	246 075	3635
Sobivad failid sisukontrolli järel	14 209	1273	0	3470
Valimisse jäävate failide arv	12 807	0	0	0

⁴ **Parser** – programme või muud teksti liigendav riistprogramm (ISO/IEC 2382)

⁵ **Pakktöötlus** (ingl *batch processing*) – Andmete töötlus üksuste kaupa (ISO/IEC 2382)

3.3.2. HTML-failidest sisu ekstraheerimine

Kuna puustruktuuriga tutvumisel ilmnis puuduvat ja üleliigset sisu, töötasin välja ekstraheerimisalgoritmi, mis arvestab varem tuvastatud puudustega ning lahendab need. Selleks tuvastasin kõik võimalikud dokumentide struktuurimallid ning nende variatsioonid. Teksti eraldamiseks on võimalik kasutada ka olemasolevaid teeke, kuid nende täpsus jääb alla taotletava sihttaseme.

Artikli sisuobjekt võib lisaks tekstile sisaldada ka lõikude vahel olevat linki muule artiklile, meediasisu (sotsiaalmeediapostitused, pildid, videod, heli) ning soovitusi edasiseks lugemiseks. Tegemist on elementidega, mille teksti puhastamiseks kasutatav tarkvara enamasti alles jätab, kuid mis vähendavad teksti representatiivsust.

Esmalt toimub puu puhastamine, mille käigus eraldatakse soovimatu meediasisu ning salvestatakse eraldi võti-väärtus paaridena lisades puusse vastava võtmega asenduselemendi. Eraldamiseks on kolm funktsiooni.

1. Esimene tuvastab ühiste tunnuste põhjal kõik lõiguvahelised viited teistele artiklitele, eraldab need puust ning lisab asenduselemendi. Teksti sisaldav vanemelement säilib.
2. Teine funktsioon eraldab manusobjektidena esineva meediasisu, mis hõlmab nii sotsiaalmeediat kui ka pilte, videoid ja helifaile. Leitavad isendid on määratud sisuanalüüsi põhjal ning võivad sisaldada ka muid otsitavaid elemente. Sõltuvuste vältimiseks töödeldakse neid funktsioonis kindlas järjekorras. Kuna leitud isendid eemaldatakse tekstist, siis ei ole neid võimalik mõne muu kriteeriumi põhjal uuesti leida.
3. Kolmas funktsioon tuvastab kõik sellised elemendid, mis sisaldavad soovitusliku lugemise loetelu ja viidet sellele. Leitakse lugemise loetelule viitav väljend, mille variandid on varasema analüüsi tulemusel eraldi muutujas defineeritud. Seejärel leitakse elemendile lähim järjestamata loetelu ning selle sisu käsitletakse soovitusliku lugemise loeteluna.

Eraldatud elementide hulgas oli kõige rohkem pilte ja videoid ($n=23275$) ning lõiguvahelisi linke teistele artiklitele ($n=14965$). Samuti eemaldasin edasise lugemise soovitude loetelud ($n=2490$). Sotsiaalmeediasisust esines Twitteri ($n=346$), Facebooki ($n=417$) ja Instagrami ($n=20$) postitusi. Muude tüüpide hulka kuulus meedia ($n=32$) ning tabelid ($n=3$).

3.3.3. Metaandmete leidmine ja lisamine

Järgmiseks leitakse dokumentide metaandmed. HTML-failist oli võimalik leida teksti pealkiri ning olemasolu korral selle alternatiivne versioon, autor, võtmesõnad, avaldamisaeg ning muutmisaeg ja teema. Töötlus hõlmab päise (*head*) ja keha (*body*) elemente ning neile rakenduvaid otsingufunktsioone.

Järgnev kirjeldus käib K1_1 kohta. Enamasti olid metaandmed päise metasildi parameetritest kättesaadavad, kuid mõnel juhul ei olnud need nõuetekohased. Sel juhul lisati väärtused nende olemasolu korral dokumendi kehast. Artikli pealkiri ja võimalikud URL-id olid kõigil juhtudel olemas. Metasildist leitav alternatiivne pealkiri oli 39% artikkelitest (n=4996). Autor oli olemas 6,43% artikkelitest (n=823) ning erinevaid autoreid oli 48. Autorit otsitakse esmalt päisest (leitud 52% juhtudest) ning selle puudumisel dokumendi kehast (leitud 48% juhtudest). Võtmesõnad olid 90,5% artikkelitest (n=11575) ning 54,3% dokumentidest olid need päisest leitavad. Teema esines ainult dokumendi kehas ning oli määratud 30,35% artikkelites (n=3887).

3.3.4. Tekstielementide töötlus

Lõpuks jäid alles vaid teksti sisaldavad elemendid. Tüübilt võisid need olla pealkirjad, alapealkirjad, sisulõigud või tsitaadid. Nende masintöötluseks sobivale kujule viimiseks kasutasin töötlusfunktsiooni, mis määras iteratiivselt neile tüübi, leidis neis sisalduvad hüperlingid ning eraldas teksti.

3.4. Tekstide puhastamine

Selles etapis kirjeldatu rakendub etappides 3.1 ja 3.2 loodud korpustele. Keskendun esmalt artiklile eelnevatele ja järgnevatele osadele, seejärel artikli teksti sees olevale.

3.4.1. Artikli põhiosast välja jääva sisu eraldamine

Artikli esimese lause mustrid

K1-IMT allikates on artiklite esimestel lausetel enamasti kindel vorm. Tuvastan esmalt kõigis allikates standardvormi ning võimalusel ekstraheerin sellest ka teabe avaldamise koha ja aja ning autori kohta. Peale vormi tuvastamist kontrollin selle esinemist artiklites. Teatud juhtudel võib mustriks vastavale esimesele lausele eelneda autori nimi või avalause, mille saab sel juhul vastavasse veergu salvestada. Selles etapis võtan nimeüksuste tuvastamiseks kasutusele ka Pythoni teegi EstNLTK-d ning korpuse sisu arvestades loodud järelmärgendaja (vt Lisa 2).

Pealkiri ja avalause

Järgmiseks kontrollin pealkirja ja avalauset. Kui need kattuvad, jääb alles ainult pealkiri. Kui need erinevad, siis määrän algoritmi (vt Lisa 1) abil, kumb on pealkiri ja kumb avalause. Kui eelmises etapis leiti eraldi avalause, siis kasutan eelmainitud algoritmi sobiva pealkirja ja avalause määramiseks.

3.4.2. Vorming

Leitakse kõik ebastandardised p -isendid. See tähendab, et p -isendi lõpus on kirjavahemärk, mis ei ole lauselõpumärk (.?!). Kategoriseerin käsitsi kõik võimalikud märgid ning määrasin neile vastavad edasised operatsioonid. Mõnel juhul piisab vormingu korrigeerimisest, kuid enamasti on lisatöötlus vajalik. Leitakse ka kõik sellised p -isendid, mille lõpus puudub kirjavahemärk.

3.4.3. Mittestandardised teksti osad

Kirjavahemärkide standardiseerimine

Tuvastan tekstis olevad ebastandardised kirjavahemärgid ja töötlen tekste vastavalt kirjavahemärgi tüübile. Samuti tuvastan *p*-isendid, millel lõpumärk puudub. Kui üks lõik on jaotunud mitmeks *p*-isendiks, siis need ühendatakse. Samuti tuvastan tsitaadid, mida hiljem eraldi märgendajaga muudele EstNLTK kihtidele lisatakse. Loendid ühendada

Teksti mittekuuluvad osad

Eelmises etapis alles jäänud mittestandardised *p*-isendid on Ko1 ja Ko2 automaatsel kogumise käigus alles jäänud osad, mis ei kuulunud artikli teksti hulka. Need hõlmavad soovitusi edasiseks lugemiseks, tekstisisesid link teistele artiklitele, piltide allkirju, sotsiaalmeediapostitusi ja küsitlusi.

Muu

Lisaks eemaldatakse võõrkeelsed lõigud. Ko2 pärit tekstide korral on võimalikud võõrkeelsed lõigud juba tekstide kogumisel tuvastatud ning need saab sisust lähtuvalt eemaldada või alles jätta. Kui artiklis on mõni *p*-isend, mis on täiesti võõrkeelne ning tegemist pole nimeüksusega, siis jääb see välja. Lisaks jäävad välja ka artiklite alguses või lõpus olevad võõrkeelsed kokkuvõtted.

Metaandmed

Metaandmete eemaldamisel leitakse ja eemaldatakse tekstist URL-id ja potentsiaalselt tundlikud andmed.

3.5. Järeltöötuse tulemused

K1_1 tekstides jäi pealkirjadest samaks 79,7%, avalausedest 55,8% ning tekstidest 58,5%, kusjuures keskmiselt jäi samaks 93,6% *p*-isenditest. K1_2 tekstides jäi pealkirjadest samaks 62,2%, avalausedest 69,9% ning tekstidest 22,4%, kusjuures keskmiselt jäi samaks 94,9% *p*-isenditest.

K1_3 tekstides pealkiri ja avalause ei muutunud. Tekstidest jäi samaks 59%, kusjuures keskmiselt jäi samaks 94,8% *p*-isenditest. K1_4 tekstides jäi pealkirjad enamjaolt samaks (93,9%) ning avalaused ei muutunud. Tekstidest jäi samaks 41%, kusjuures keskmiselt jäi samaks 90% *p*-isenditest.

4. Märgendamine ja järeltöötlus

Korpuse sisu tähistustasandil analüüsimiseks on vaja see morfoloogiliselt märkendada. See võimaldab hinnata tekstide ja lausete pikkust, sõnaliikide jaotust ning luua sõnade sagedusloendeid. Siinse töö kontekstis on oluline roll ka nimeüksustel, mille jaotust ja liike saab samuti analüüsida.

Kasutasin tekstide automaattöötluse jaoks Pythoni teekide kogu EstNLTK (Laur jt 2020). Lisasin puhastatud tekstidele alapeatükis 2.3.6. kirjeldatud vaikesätetega märkenduskihid. Märkendaja tulemuse korrigeerimiseks andsin initsialiseerimisel lisamärkendajana kaasa UserDictTaggeri, mis andis automaatanalüsaatorile teada, milline peab olema kindlate mitmese või vale analüüsiga sõnade õige märkend.

4.1.1. EstNLTK märkenduskihid

EstNLTK teisendab lisamiseks sisendteksti Text objektiks ning lisab sellele sisendparameetritega määratud Layer-tüüpi kihid. Vaikimisi lisatakse neli segmenteeritud teksti kihti (*tokens*, *compound_tokens*, *words*, *sentences*) ning morfoloogilise analüüsi kiht (*morph_analysis*).

Esiteks lisatakse sõnade kiht (*tokens*). Selleks tükeldatakse tekst sõnedeks, määraes kirjavahemärkide või tühikute põhjal iga sõne alguse ja lõpu. Järgmiseks lisatakse mitmest osast koosnevate sõnade kiht (*compound_tokens*). Järgmiseks lisatakse sõnade kiht (*words*). Selleks tükeldatakse tekst sõnadeks lähtudes sõnade ja mitmest osast koosnevate sõnade kihtidest. Kihi atribuudiks on sõna normaliseeritud kuju (*normalized_form*) ning tõeväärtus (*ambiguous*) vastamaks sellele, kas sõnal on mitu normaliseeritud kuju. Loodud kihtide põhjal lisatakse morfoloogilise analüüsi kiht. See põhineb sõnade ja lausete kihtidel, analüüs lisatakse igale sõnade kihi elemendile. (EstNLTK dokumentatsioon 2026)

Morfoloogilise analüüsi kihil on kaheksa atribuuti: normaliseeritud tekst (*normalized_text*), algvorm (*lemma*), sõnatüvi (*root*), sõnatüve sõned (*root_tokens*), sõnalõpp (*ending*), kliitik (*clitic*), vorm (*form*) ja sõnaliik (*partofspeech*). (EstNLTK dokumentatsioon 2026e; EstNLTK dokumentatsioon 2026)

4.1.2. Mitmese analüüsiga sõnade ühestamine

Morfoloogilise automaatanalüüsi väljund võib sisaldada mitmese analüüsiga elemente. See tähendab, et analüsaator ei ole sõna võimalike vormide hulgast õiget leidnud ning annab väljundiks kõik võimalikud. Mitmeste analüüsides hulga vähendamiseks saab lähtuda kindlatest reeglitest, mille põhjal võimalike vormide hulgast õige määrata. Töötluse eesmärgiks on jätta alles üks analüüs. Morfoloogilise märgenduse kihi esmane ühestamine kujutab endast mitmese analüüsiga sõnade kõige tõenäolisema märgendi alles jätmist.

See kehtib näiteks juhtudel, mil automaatanalüüsi käigus lisatakse sõnale nii konjunktsiooni kui adverbialanalüüs. Siis jäetakse alles ainult konjunktsiooni (näiteks *kui*, *nagu*) analüüs. Sagedaste mitmeste analüüsides korral on võimalik morfoloogilisele märgendajale initsialiseerimisel selliste sõnade õiged märgendid anda. Tegusõna mineviku partitsiipidele (*nud/tud*) lisab EstNLTK automaatmärgendaja vaikimisi neli analüüsi, mille hulgast jäetakse kokkuleppeliselt alati alles tegusõna mineviku partitsiibi tõlgendus.

Sagedasemad sõnaliikide vahelised mitmesused olid adverb vs konjunktsioon, ordinaalnumeraal vs pronoomen (näiteks *teine*), kardinaalnumeraal vs pronoomen (näiteks *üks*). Mitmese analüüsi, aga sama sõnaliigiga sõnade puhul esines probleeme verbi ja pronoomeni ainsuse ja mitmuse vormi eristamisel.

Kõigi mitmeste analüüsides korral, mida kokkuleppelistest reeglitest lähtudes automaatselt lahendada ei saa, jäetakse alles esimene analüüs.

4.1.3. Nimeüksuste tuvastamine ja järeltöötlus

Kasutasin nimeüksuste tuvastamiseks EstNLTK närvivõrgupõhist nimeüksuste tuvastajat EstBERTNERTagger (Sirts). Mudel tuvastab isikunimed (PER), kohanimed (LOC), organisatsioon (ORG), geopoliitilised üksused (GPE), valuuta (MONEY), osakaalud (PERCENT) tooted (PROD), nimetused (TITLE), kuupäevad (DATE), aja (TIME) ja sündmused (EVENT).

Uudistekstides viidatakse palju teistele uudisteagentuuridele ning sotsiaalmeediapostitustele, kuid ükski olemasolevatest nimeüksuste kategooriatest ei kohaldunud kummalegi neist täielikult. Sellest tulenevalt lisasin töötuse käigus kategooriad uudiste allikad (NEWS) ja sotsiaalmeediaplatvormid (SOCIAL_MEDIA). Sotsiaalmeedia hõlmas sotsiaalsõrgustike nimesid (nt Facebook, Twitter, Instagram).

Peale vaikesätetega nimeüksuste märgenduskihi lisamist kasutasin spetsiaalselt korpuse jaoks loodud järelmärgendajat (vt Lisa 2), mis tuvastab morfoloogilise analüüsi kihis määratud sõnaliikide põhjal nimeüksused ning leiab nende vasted NER-kihis. Vaste puudumisel lisatakse see kihti ning olemasolu korral muudetakse NER-kihti vaste liigi alusel.

5. Esialgne kirjeldav statistika

Tabel 3. Tekstide arv ja lõplikud suurused allkorpustes

Kategooria	Allikas	Tekstide arv korpuses				Sõned	Sõnad	Laused
		Ko1	Ko2	Ko3	kokku			
K1-IMT	K1_1	3509	1879	12 807	18 195	6 847 507	6 621 962	366 787
	K1_2	54	89	0	143	131 512	129 772	8530
	K1_3	0	61	0	61	36 768	36 059	2164
	K1_4	13	199	0	212	67 352	65 224	3639
K1-IMT kokku		3576	2228	12 807	18 611	7 083 139	6 853 017	381 120

Tabel 4. Tekstide arv ja suurused võrdluskorpustes

Kategooria	Allikas	Tekstide arv korpuses			Sõned	Sõnad	Laused
		Ko1	Ko2	Ko3			
K2-ALT	K2_1	10 378			1 030 540	838 834	48 483
	K2_2	22 089			3 905 205	3 178 737	183 723
	K2_3	3811			493575	401 757	23 221
K2-ALT kokku		36 278			5 423 896	4 414 912	255 170
K3-ERR	K3_1	202 187			56 829 919	46 258 102	3 827 558

K1_1 tekstide keskmine pikkus on 364 sõna ja 20 lauset, K1_2 tekstide keskmine pikkus 907 sõna ja 60 lauset, K1_3 tekstide keskmine pikkus 907 sõna ja 35 lauset ning K1_4 tekstide keskmine pikkus 591 sõna ja 17 lauset.

5.1. Sõnaliikide jaotus

		Suhteline sagedus (mln sõna kohta)			Suhteline levik protsentides		
		K1	K2	K3	K1	K2	K3
		(IMT)	(ALT)	(ERR)	(IMT)	(ALT)	(ERR)
substantiiv		257530	251573	256603	98,7%	100%	100%
verb		138177	168273	165927	98,7%	99,9%	99,6%
pärisnimi		81729	62009	80907	98,3%	99%	96,9%
adjektiiv	käänduv	19646	22988	20982	58,6%	57,7%	52,2%
	käändumatu	2087	2225	1909	32,8%	34,5%	24,9%
adverb		63063	92038	84568	97,4%	97,9%	95%
pronoomen		57331	82109	71310	96,7%	96,7%	92%
konjunktsioon		53683	71074	64765	97,9%	98,4%	95,5%
adpositsioon		20176	21411	20459	95,1%	93,2%	85,3%
numeraal	kardinaal	16255	11371	21488	85,6%	74,1%	78,4%
	ordinaal	11146	4413	8296	94%	55,4%	61,5%
lühend		11146	9485	8687	83%	78,7%	60%
interjektsioon		191	527	257	5,2%	13,2%	3,8%
kirjavahemärk		144979	154332	151717	98,7%	99,8%	99,7%

Tabel 6. Sõnaliikide suhtelised sagedused

Tabeli põhjal ilmneb, et infomõjutustekstides esineb verbe suhtelise sageduse järgi arvestades miljoni sõna kohta alternatiivmeedia tekstidega võrreldes 30 096 ning ERR tekstidega võrreldes 27 750 vähem. Pärisnimesid esineb infomõjutustekstides miljoni sõna kohta kõige rohkem ning erinevus alternatiivmeediatekstidega on märgatav (K1-s 19 720 rohkem), kuid erinevus ERR-iga pole eriti suur (K1-s 822 rohkem). Adjektiive on suhtelise sageduse järgi mõlemast võrdluskorpusest vähem (K2-s 3342 rohkem, K3-s 1336 rohkem), kuid suhtelise leviku poolest kõige rohkem. Adverbe ja pronomeneid on võrdluskorpustes oluliselt rohkem (adverbe 25 240, pronomeneid 19 379). Lühendeid on võrdluskorpustest keskmiselt 2060 rohkem.

Suhteliste sageduste ja suhtelise leviku võrdluse põhjal saab järeldada, et infomõjutustekstides on alternatiivmeedia ja ERR-iga võrreldes rohkem pärisnimesid ja lühendeid, oluliselt vähem verbe, adjektiive, adverbe ja pronomeneid.

6. Kokkuvõte

Siinse töö põhieesmärk oli luua infomõjutustekstide korpus, mis võimaldaks edasist analüüsi. Korpuse loomise käigus tekkinud lisaeesmärkideks olid korpuse kvaliteedi tõstmine ning tsitaatide ja nimeüksuste tuvastamiseks korpuse eripäradega arvestavate märgendajate loomine. Töö praktiline osa sisaldab kahel erineval viisil korpuse loomiseks kasutatavaid töövooge, lisaks ka funktsioone tekstide töötlemiseks ja nendest teabe ekstraheerimiseks. Nimetatud eesmärgid said täidetud ning töö tulemusel valmis korpus suurusega 6 853 017 sõna.

Esmaste tulemuste põhjal saab eeldada, et infomõjutustekstide korpuse koostamine oli põhjendatud. Lisaeesmärgiks oli esmase kirjeldava statistika loomine, mille tulemused toetavad töö alguses püstitatud nullhüpoteesi, et informatsioonilise mõjutustegevuse keelekasutus erineb muul eesmärgil loodud meediatekstide omast. Hüpoteesi ei ole olemasolevate andmete põhjal võimalik veel vastu võtta ega ümber lükata, kuna see eeldab põhjalikumat sisuanalüüsi.

Töö tulemusena valminud korpus avab võimaluse leida infomõjutustegevuse tekstide keelekasutuse eripärad. Edasised võimalikud töösuunad hõlmavad võtmesõnapõhist ja mitmetasandilist analüüsi ning sündmuste ja narratiivide tuvastamist.

6.1. Piirangute kirjeldus

Esialgne töökava nägi ette olemasoleva korpuse materjalidest allkorpuse loomist, selle käsitsi temaatilist märgendamist ja analüüsi. Töö käigus ilmnis, et olemasolevas korpuses sisalduvad tekstid ei pruugi vastata nende originaalkujule avaldamiskohas. Samuti tuli välja, et tekstid vajavad täiendavat töötlust ja puhastamist, kuna need sisaldasid teksti mittekuuluvaid osi. Sellest tulenevalt muutus töö iseloom ning uueks eesmärgiks sai korpuse kvaliteedi tõstmine. Esinduslikkuse tagamiseks lõin veebitekstidest uue korpuse, mille tekstid oli olemasolevas korpuses sisalduvate vastetega võrreldes pikemad.

Probleemiks oli ka osade tekstide algkuju kättesaadavus. Nimelt oli tekstide sisu kontrollimisel võimalik võrrelda neid ainult arhiveeritud veebiversioonidega, mille puudumisel polnud vastavust originaalkujule võimalik hinnata.

7. Kirjandus

- Aïmeur, Esma, Sabine Amri & Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining* 13(1). 30. <https://doi.org/10.1007/s13278-023-01028-5>.
- Alyukov, Maxim, Maria Kunilovskaya & Andrei Semenov. 2023. Wartime Media Monitor (WarMM-2022): A Study of Information Manipulation on Russian Social Media during the Russia-Ukraine War. *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 152–161. Dubrovnik, Croatia: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2023.latechclfl-1.17>.
- Barbu, Eduard, Somnath Banerjee, Marija Isupova & Yukai Zeng. 2024. *Narrative Detection and Topic Modelling in the Baltics*. NATO Strategic Communications Centre of Excellence.
- Biber, Douglas. 1993. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4).
- Burchell, Laurie, Ona de Gibert, Nikolay Arefyev, Mikko Aulamo, Marta Bañón, Pinzhen Chen, Mariia Fedorova, jt. 2025. An Expanded Massive Multilingual Dataset for High-Performance Language Technologies (HPLT). arXiv. <https://doi.org/10.48550/arXiv.2503.10267>.
- Conroy, Nadia K., Victoria L. Rubin & Yimin Chen. 2015. Automatic deception detection: Methods for finding fake news. *Proceedings of the Association for Information Science and Technology* 52(1). 6–10. <https://doi.org/10.1002/pra2.2015.145052010082>.
- Fagan, Craig, Andres Jõesaar & Andres Kõnno. 2020. *Väärinforiskid Eesti Meediaturul*. The Global Disinformation Index (GDI - Globaalne Väärinformatsiooni Indeks).
- Kaitsepolitseiamet. 2014. Kaitsepolitseiameti aastaraamat 2014.
- Kaitsepolitseiamet. 2025. Kaitsepolitseiameti aastaraamat 2024/2025.

- Kerge, Krista. 2002. Aja- ja ilukirjandusteksti süntaktilise keerukuse dünaamika XX sajandil. *TPÜ eesti filoloogia osakonna veebitoimetised* 1.
- Lange-Ionatamishvili, Elina & Kamil Mikulski. 2024. What Are ‘Information Influence Operations’ and Where Do They Fit Conceptually?. Ahonen, Anneli, Arsalan Bilal, Gunhild Hoogensen Gjørsv, Juris Jurāns, Martin Kragh, Kristofers Kārlis Krūmiņš, Elina Lange-Ionatamishvili, jt. *Russia’s Information Influence Operations in the Nordic – Baltic Region*, 7–14. NATO Strategic Communications Centre of Excellence.
- Lazer, David M. J., Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, jt. 2018. The science of fake news. *Science* 359(6380). 1094–1096. <https://doi.org/10.1126/science.aao2998>.
- Laur, Sven, Siim Orasmaa, Dage Särge & Paul Tammo. 2020. EstNLTK 1.6: Remastered Estonian NLP Pipeline.
- Modzelewski, Arkadiusz, Giovanni Da San Martino, Pavel Savov, Magdalena Anna Wilczyńska & Adam Wierzbicki. 2024. MIPD: Exploring Manipulation and Intention In a Novel Corpus of Polish Disinformation. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 19769–19785. Miami, Florida, USA: Association for Computational Linguistics. <https://doi.org/10.18653/v1/2024.emnlp-main.1103>.
- Oshikawa, Ray, Jing Qian & William Yang Wang. 2020. A Survey on Natural Language Processing for Fake News Detection. arXiv. <https://doi.org/10.48550/arXiv.1811.00770>.
- Pamment, James & Victoria Smith. 2022. ATTRIBUTING INFORMATION INFLUENCE OPERATIONS.
- Paquot, Magali & Stefan Th. Gries (toim). 2020. *A Practical Handbook of Corpus Linguistics*. Cham: Springer International Publishing. <https://doi.org/10.1007/978-3-030-46216-1>.
- Pilvik, Maarja-Liisa, Kadri Muischnek, Liina Lindström, Ann Veismann, Jane Klavan, Jelena Kallas, Kristina Koppel, jt. 2025. Korpused lingvistilises uurimistöös. Praktiline käsiraamat.

- Potthast, Martin, Johannes Kiesel, Kevin Reinartz, Janek Bevendorff & Benno Stein. 2018. A Stylometric Inquiry into Hyperpartisan and Fake News. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 231–240. Melbourne, Australia: Association for Computational Linguistics. <https://doi.org/10.18653/v1/P18-1022>.
- Richardson, Leonard. 2025. BeautifulSoup 4. <https://pypi.org/project/beautifulsoup4/>.
- Rizgelienė, Ieva, Vilma Zubaitienė, Nerijus Maliukevičius & Virginijus Marcinkevičius. 2025. HALT-PROP: Human-Annotated Lithuanian Textual Corpus for Propaganda Narratives and Techniques. *Sci Data*. <https://doi.org/10.18279/MIDAS.260057>.
- Rubin, Victoria L., Yimin Chen & Nadia K. Conroy. 2015. Deception detection for news: Three types of fakes. *Proceedings of the Association for Information Science and Technology* 52(1). 1–4. <https://doi.org/10.1002/pra2.2015.145052010083>.
- Schütz, Mina, Daniela Pisoiu, Daria Liakhovets, Alexander Schindler & Melanie Siegel. 2024. GerDISDETECT: A German Multilabel Dataset for Disinformation Detection.
- Sirts, Kairit. Estonian Named Entity Recognition: New Datasets and Models.
- Solopova, Veronika, Viktoriia Herman, Christoph Benz Müller & Tim Landgraf. 2024. Check News in One Click: NLP-Empowered Pro-Kremlin Propaganda Detection.
- Solopova, Veronika, Oana-Iuliana Popescu, Christoph Benz Müller & Tim Landgraf. 2023. Automated Multilingual Detection of Pro-Kremlin Propaganda in Newspapers and Telegram Posts. *Datenbank-Spektrum* 23(1). 5–14. <https://doi.org/10.1007/s13222-023-00437-2>.
- Vabariigi Valitsus. 2023. Eesti julgeolekupoliitika alused 2023.
- Ventsel, A & M-L Madisson. 2024. POLIITIKASEMIOOTIKA VAADE VANDENÕUTEOORiatele. *EESTI VABARIIGI PREEMIAD* 28(1). 224. <https://doi.org/10.3176/evp.2024.11>.
- Voog, Aivar, Jaanika Hämmal, Kaisa Esko, Külliki Seppel, Marju Lauristin, Marre Karu & Triin Vihalemm. 2024. *Eesti ühiskonna lõimumismonitooring 2023*. Kultuuriministeerium. <https://www.kul.ee/EIM2023>.

Yu, Seunghak, Giovanni Da San Martino, Mitra Mohtarami, James Glass & Preslav Nakov.
2021. Interpretable Propaganda Detection in News Articles. *Proceedings of the Conference Recent Advances in Natural Language Processing - Deep Learning for Natural Language Processing Methods and Applications*, 1597–1605. INCOMA Ltd. Shoumen, BULGARIA. <https://doi.org/10.26615/978-954-452-072-4> 179.

väärinfo. *Militerm*. <https://sonaveeb.ee/search/unif/dlall/mil/väärinfo/1/est>.

desinformatsioon. *Militerm*. <https://sonaveeb.ee/search/unif/dlall/mil/desinformatsioon/1/est>.

kuriinfo. *Militerm*. <https://sonaveeb.ee/search/unif/dlall/mil/kuriinfo/1/est>.

informatsiooniline	mõjutustegevus.	<i>Militerm.</i>
https://sonaveeb.ee/search/unif/dlall/mil/informatsiooniline%20mõjutustegevus/1/est .		

EVS-ISO/IEC 2382. Parser.

EstNLTK dokumentatsioon.

8. Creating a corpus of media texts used in information influence activities.

Summary

The aim of this bachelor's thesis is to establish a basis for the future computational analysis of the linguistic characteristics of information influence activities conducted through media. To achieve this, a corpus consisting of 18 611 Estonian media texts published, between 2014-2022 and originating from four news outlets, was created. In addition, a preliminary descriptive statistical analysis of the corpus was performed.

Lisad

Lisa 1. Tekstitöötlusalgoritmid

Kui pealkirja ja avalauset etteantud kirjes pole, siis püütakse need esmalt leida. Seejärel leitakse viie erineva kombinatsiooni Levenshteini kaugused. Kontrollitakse URL-ist parsitud pealkirja ja tavalise pealkirja sarnasust, avalause ja pealkirja sarnasust, pealkirja ja esimese p -isendi sarnasust, URL-ist parsitud pealkirja ja avalause sarnasust. Samuti leitakse pärsitud pealkirja ühiste sõnade arv pealkirja ja avalausega.

Kui tekstist leiti uus avalause, siis määratakse see pealkirjaks või avalauseks vastavalt sõnede sarnasustele. Enamasti on pealkiri selle eeldatavale originaalkujule sarnasem kui avalause ning avalause määratakse alternatiivseks avalauseks. Muul juhul määratakse avalause pealkirjaks.

Kui tekstist uut avalauset ei leitud, siis kontrollitakse vastavust veel kolmel juhul. Kui esialgse pealkirja ja avalause sarnasuse määr on vähemalt 90%, siis jäetakse neist alles see, mis on pealkirja eeldatavale originaalkujule sarnasem. Kui esialgne pealkiri on eeldatava originaalkujuga sama, siis määratakse esialgne avalause uueks avalauseks.

Kui esialgse avalause ja pealkirja eeldatava originaalkuju sarnasuse määr on vähemalt 80% ning see on suurem kui pealkirja ja selle eeldatava originaalkuju sarnasuse määr

VÕI kui esialgse avalause ja pealkirja eeldatava originaalkuju ühisosa (pikkuse mõttes) sarnasuse määr on vähemalt 80%

VÕI kui esialgse avalause ja pealkirja eeldatava originaalkuju ühisosa (sõnade mõttes) sarnasuse määr on vähemalt 90% ning see on suurem kui pealkirja ja selle eeldatava originaalkuju ühisosa (sõnade mõttes) sarnasuse määr, siis määratakse esialgne avalause uueks pealkirjaks.

Lisa 2. Nimeüksuste märgenduskihi järelmärgendaja NERPostTagger

EstNLTK nimeüksuste märgendaja ei tuvastanud kõiki tekstides olevaid nimeüksusi, kuna polnud selleks treenitud. Nimeüksuste märgenduskihi korrigeerimiseks ja täiendamiseks on loodud eraldi klass NERPostTagger, mille tööpõhimõtted sarnanevad EstNLTK märgendajatele. Märgendaja tuvastab morfoloogilise märgenduse kihist sõnaliikide põhjal kõik võimalikud nimeüksused. Seejärel kontrollib märgendaja nimeüksuse olemasolu NER-kihis.

Märgenduskihi töötlus lähtub sellest, kas ja millisel määral vaadeldaval nimeüksusel NER-kihis vaste leidub:

- Kui nimeüksusel on täielik vaste, siis muudetakse teatud juhtudel märgendi tüüpi, kuid enamasti jääb NER-i liik samaks.
- Kui vaste puudub täielikult, siis lisatakse see NER-kihti koos sobiva märgendiga.
- Kui vaste läheb üle sõnapiiride, siis kontrollitakse olemasolevate märgendite kategooriatest lähtuvalt, kas tegemist on mitme üksteisele järgneva pärisnimega või tuleb need eraldada.
- Osaliselt puuduva vaste korral töödeldakse seda vastavalt varem määratud tingimustele. Sel juhul asendatakse osaline vaste leitud nimeüksusega.

Lisafunktsioonid võimaldavad NER-kihi üksusi kombineerida ja tükeldada, otsida vasteid etteantud mustrite järgi (näiteks neli üksteisele järgnevat lemmat ja leiab kõik sellele vastavad väljendid)

ⁱ Keelelise selguse ja ühetähenduslikkuse tagamiseks pärinevad informatsiooni tüüpide eestikeelsed definitsioonid sõjandusterminoloogia töörühmalt (21774 (R-011) „Riigikaitseterminoloogia arendamine“, 2020-2024, Hendrikson jt.) ning on kooskõlas viidatavas raamistikus toodud definitsioonidega.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Marielin Pomberg-Kalm,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Infomõjutustekstide korpuse loomine“, mille juhendaja on Kadri Muischnek, PhD, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;
2. annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commons'i 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.

Marielin Pomberg-Kalm

28.01.2026