

Hankija: Sihtasutus Eesti Teadusagentuur
HANKELEPING nr 1.2-5/24/16



Kontaktisik:
Indrek Heinla
indrek.heinla@etag.ee
Tel 552 4134

SA Eesti Teadusagentuuri tööprotsesside automatiseerimise võimalikkuse analüüs

Aruanne

Koostasid:
Birgitta Ojamaa, Ines Anett Nigol, Ingvar Baranin, Silver Traat, Kalev Koppel, Raul Sirel,
Marit Asula

Uuringut rahastas SA Eesti Teadusagentuur Regionaalarengu Fondist toetatava sekkumise „Teadus-, arendus- ja innovatsioonitegevuste tulemuste rakendamise võimekuse tõstmine ühiskonnas ning selleks soodsa poliitikakeskkonna loomine“ (RITA+ alategevuse 4) kaudu

Kuupäev: 26.08.2024

Sisukord

Mõisted ja definitsioonid.....	4
1. Sissejuhatus.....	5
1.1. Tegevused ja tulemused.....	5
1.1.1. Teadusvaldkonna määramine rahastustaotlustele ja publikatsioonidele.....	5
1.1.2. Arengukavade vastavusanalüüs.....	5
1.1.3. Publikatsioonidest eestikeelsete populaarteaduslike kokkuvõtete kirjutamine.....	6
1.1.4. Eesti Teadusinfosüsteemi tekstiotsingute parandamine.....	6
1.1.5. Taotlustele sobivate hindajate leidmine.....	6
2. Teadusagentuuri tööprotsesside ülevaade.....	8
3. Sisendandmete analüüs.....	11
3.1 Sisendandmete kättesaamine ja töötlemine.....	11
3.2 Sisendandmete kvaliteedi hindamine.....	12
3.2.1 Sisendandmete hindamismeetodid.....	13
3.3 Sisendandmete kvaliteedi hindamise tulemus.....	14
3.3.1 Sisendandmete kvaliteedi hindamise üldtulemused.....	14
3.3.2 Publikatsioonide tüüpide hindamine.....	15
3.3.3 Autorite nimede hindamine.....	16
3.3.4 Sisukokkuvõtete hindamine.....	16
3.4 Hinnang sisendandmete kvaliteedile.....	17
4. Valdkondlike andmete analüüs.....	19
4.1 Valdkondlike andmete kvaliteedi hindamine.....	19
4.2 Valdkondlike andmete kvaliteedi hindamine.....	21
4.2.1 CERCS valdkonnad ja OpenAlexi alamvaldkonnad.....	22
4.2.1.1 Automaatne võrdlus.....	22
4.2.1.2 Keelevaldkonna lühikokkuvõtete võrdlus.....	24
4.2.2 Märksõnad ning OpenAlex alamvaldkonnad.....	24
4.3 Lõpphinnang valdkondlikele andmetele.....	25
5. Taotlustele valdkondade määramise analüüs ja prototüüp.....	26
5.1 Alusandmed ja definitsioonid.....	26
5.2 Andmete töötlus.....	26
5.3 Meetodite võrdlus.....	27
5.4 Meetodite analüüs, eksperimendid ja tulemused.....	29
5.4.1 BERT-klassifitseerija abil valdkondade määramine.....	29
5.4.2 Tagger Group mudelite abil valdkondade määramine.....	30
5.4.2.1 OpenAlex põhivaldkonna pealt treenitud Tagger Group.....	32
5.4.2.2 Ühe valdkonna määramine Tagger Group abil suurema andmestiku põhjal.....	32
5.4.2.3 Kümne valdkonna määramine Tagger Group abil suurema andmestiku põhjal...	33
5.4.3 LLM-i abil valdkondade määramine.....	36
5.4.3.1 Teiste päringute mõju LLM-i määratud valdkondadele.....	37

5.5 Valdonna määramise tulemuste kokkuvõte ja kontseptuaalne prototüüp.....	39
5.6 Lisategevused tulemuste parandamiseks.....	39
6. Arengukavade vastavusanalüüsi abilise analüüs.....	41
6.1 Alusandmed ning nende töötlemine.....	41
6.2 Arengukavadele vastavuse leidmise meetodid.....	42
6.2.1 Andmete viimine vektorkujule.....	42
6.2.2 Sarnasusotsingute liigid.....	42
6.2.3 Sisendid otsingutele.....	43
6.3. Arengukavadele vastavuse leidmise tulemused.....	45
6.3.1 Taotluste annotatsioonide sarnasusotsingute tulemused.....	46
6.3.2 Projektide annotatsioonide sarnasusotsingute tulemused.....	47
6.3.3 Sünteetiliste näidete sarnasusotsingute tulemused.....	48
6.4. Tulemuste kokkuvõte.....	48
6.5. Soovitused tulemuste parandamiseks.....	49
7. Publikatsioonist populaarteadusliku kokkuvõtte tegemine.....	51
7.1. Alusandmed.....	51
7.2. Meetodid ja mudelite valik.....	52
7.2.1 Päringud.....	52
7.2.2 Hinna arvestamine.....	53
7.3. Tulemused.....	53
7.3.1 Genereeritud sisukokkuvõtete kvaliteet.....	53
7.3.2 Juhenditest kinnipidamine.....	54
7.3.3 Mudelite hind.....	55
7.3.4 Limiidid, torked, hallutsineerimine ning muud märkused.....	57
7.4 Tulemuste kokkuvõte.....	57
7.5 Soovitused tulemuste parandamiseks.....	58
8. Eesti Teadusinfosüsteemi tekstiotsingute täiendamine.....	60
8.1 Võimalikud meetodid probleemi lahendamiseks.....	60
9. Taotlustele sobivate hindajate leidmine ja huvide konfliktiga hindaja välistamine.....	62
9.1 Annotatsiooni ja valdkonna põhjal sarnaste publikatsioonide leidmise katse.....	62
10. Soovituslikud ülesanded ja mahuhinnangud analüüsi põhjal.....	64
11. Lisad.....	67
Lisa 1. LLM-i abil valdkondade määramine, esimene päring.....	67
Lisa 2. LLM-i abil valdkondade määramine, teine päring.....	70
Lisa 3. LLM-i abil valdkondade määramine, kolmas päring.....	71
Lisa 4. Tagger Group API juhend.....	73
Lisa 5. Kokkuvõtete koostamise eksperimendis kasutatud publikatsioonide viited.....	75
Lisa 6. LLM päringu publikatsiooni annotatsioon ning sellele loodud kokkuvõte.....	76
Lisa 7. Projekti annotatsioonile ja valdkonnale sarnaseimad publikatsioonid.....	79

Mõisted ja definitsioonid

BERT mudel	BERT on Transformer-mudel, mida on võimalik rakendada erinevate loomuliku keele töötluste ülesannete lahendamiseks, sh klassifitseerimiseks.
ElasticSearch	Andmebaasisüsteem, mis võimaldab indekseerimist ning otsingut.
generatiivne suur keelemudel (LLM)	Generatiivsed suured keelemudelid (edaspidi LLM) on Transformer-mudelid, mis genereerivad sisendi põhjal teksti.
lemmatiseerimine	Tekstis sisalduvate sõnade viimine algvormi, nt <i>karude</i> lemma on <i>karu</i> .
regulaaravaldis	Regulaaravaldis on otsingumuster, mille abil saab tekstilise info seast otsida, teatud sümboleid, tähemärke ning sõnade-fraaside erinevaid variante.
sarnasusotsing	Otsing sisulise sarnasuse põhjal, mille alamliik on vektorotsing, kus tekst, mille hulgast otsitakse, on viidud vektorkujule.
segmentimine	Dokumendi väiksemateks üksusteks jagamine, näiteks lehekülgedeks.
sõne	Kirjatähtede jada või kõneühik, mis võib, kuid ei pruugi kattuda sõnaga.
Tagger Group	Tagger on statistiline mudel, mis kasutab logistilist regressiooni ning tugivektorklassifitseerijat, et sõna- ja tähemärgivektorite abil iga klassi (valdkonna) tunnused määrata, näiteks sõnade või fraaside esinemise tekstis. Tagger Group hõlmab mitme Taggeri treenimist korraga, üks tekst võib seostuda mitme valdkonnaga.
vektoriseerimine	Teksti vektorkujule viimine, et sooritada vektorotsinguid.

Kvaliteedimõõdikud:

Kõik tulemuste hindamiseks kasutatud mõõdikud on arvatud skaalal 0.00–1.00, mis vastab sisuliselt protsendile, ehk siis 1.00 puhul on mudeli antud mõõdiku kvaliteet kõige parem (100%) ja 0.00 puhul on mudeli antud mõõdiku kvaliteet kõige halvem (0%).

Täpsus (*precision*) – näitab õigete positiivsete arvu kõigi positiivsete tulemuste seas. Ehk siis kui meil on näiteks ühe valdkonna puhul kaks valepositiivi ja kaks õiget positiivi, siis oleks täpsus 0.5 (ehk 50%).

Saagis (*recall*) – näitab kui palju õigeid positiivseid tulemusi leiti kõigi tulemuste seast, ehk näitab vale-negatiivide arvu. Näiteks kui leiti kaks õiget positiivset, aga veel kahte ei leitud, siis saagis on 0.5 (ehk 50%).

F1 – Täpsuse ja saagise harmooniline keskmine, mis näitab keskmist andmete kvaliteeti.

Kõik kvaliteedimõõdikud on [scikit.learn](https://scikit-learn.org/) Python teegist, kasutades erinevaid keskmise arvutamise algoritme.

1. Sissejuhatus

Projekti eesmärgiks oli analüüsida SA Eesti Teadusagentuuri (ETAG) viie tööprotsessi optimeerimise võimalikkust tehisintellekti või muude uudsete tehnoloogiliste lahenduste abil.

Projekti raames intervjueriti ETAGi juhtkonda ja töötajaid, et täpsustada probleemide skooopi, analüüsiti sisendandmete kvaliteeti ning eksperimenteeriti erinevate masinõppe mudelite ja algoritmidega, et pakkuda välja tööprotsesside optimeerimise edasine suund ning anda ressurssihinnang selleks vajalike tööriistade loomisele või rakendamisele.

1.1. Tegevused ja tulemused

1.1.1. Teadusvaldkonna määramine rahastustaotlustele ja publikatsioonidele

Probleem: 1) ETAG hindab rahastustaotlusi valdkondlike ekspertkomisjonide abil. Taotlusele õige valdkonna omistamine tagab, et taotlus läheb hindamisele kõige kompetentsemale ekspertkomisjonile. 2) ETISes sisalduvatele publikatsioonidele teadusvaldkonna omistamine võimaldab paremini saada ülevaadet Eesti teaduse olukorrast.

Tegevused: Analüüsi sisendandmeteks olid Eesti Teadusinfosüsteemi (ETIS) projektide eestikeelsed annotatsioonid, OpenAlex andmebaasi publikatsioonide ingliskeelsed annotatsioonid (kokku üle 160 tuh annotatsiooni) ning väikeses koguses ka ETISe rahastustaotluste teadusliku põhjenduse tekstid (10 tk). Sisendandmete põhjal treeniti kahte liiki masinõppemudeleid (BERT ja Tagger Group). Lisaks katsetati ka LLM-mudeleid ChatGPT ja OpenAI GPT-4o.

Tulem: Parima tulemuse andsid Tagger Group mudelid ETISe projektide eestikeelsete annotatsioonide põhjal. Taotlustele määras masinõppemudel õige valdkonna 81% juhtudest. Tulemust saaks veelgi parandada, ent see eeldab lisategevusi sisendandmete (käsitsi) puhastamisel ja mudelite korrigeerimisel. Teiste meetodite tulemus jäi alla 60%.

1.1.2. Arengukavade vastavusanalüüs

Probleem: Rahastustaotluste menetlemisel peab ETAG sageli hindama, kas taotlus vastab mõne arengukava eesmärkidele. Tegevus on ajamahukas ja veaohklik.

Tegevused: Eksperimenteeriti kahe erineva sarnasusotsingu liigiga (MLT ja KNN). Vasteid arengukavadele otsiti rahastustaotlustest, ETISe projektide annotatsioonidest ja sünteetilistest näidetest.

Tulem: Leiti, et taotluse ja arengukava vastavuse otsing oli tulemuslik, kui teksti polnud radikaalselt muudetud. Parima tulemuse andis vektorotsing (KNN). Rohkem muudetud tekstidele vaste leidmine vajab täiendavat eksperimenteerimist.

1.1.3. Publikatsioonidest eestikeelsete populaarteaduslike kokkuvõtete kirjutamine

Probleem: Teadlaskond peaks oma tegevusi laiemale avalikkusele paremini tutvustama, samas on teaduspublikatsioonidest populaarteaduslike kokkuvõtete kirjutamine ajamahukas ja teadlastele ka harjumatu tegevus.

Tegevused: Sisendiks olid seitsme teaduspublikatsiooni terviktekstid. Kokkuvõtte loomise töövahendina võrreldi viit suurt generatiivset keelemudelit: GPT-4-turbo, GPT-3.5-turbo, Claude Sonnet 3.5, Claude Opus 3 ning Gemini Pro 1.5.

Tulem: Enamik keelemudeleid said ülesandega hakkama, ehkki kõigil olid omad tugevused ja nõrkused (teksti pikkuse limiidid, hallutsineerimine, hind). Tulemuses esines väiksemaid õigekeelsusvigu ning genereeritud tekstid vajavad toimetamist.

1.1.4. Eesti Teadusinfosüsteemi tekstiotsingute parandamine

Probleem: ETISE otsingud tuleks muuta paindlikumaks, et need leiaks vastuseid ka sõltumata otsisõna käändelõppude erinevusest või väikestest ebatäpsustest.

Tulem: Analüüsi tulemusena soovitame probleemi lahendamiseks katsetada sarnasusotsinguid (sarnane lahendus nagu punktis 1.2). Käändelõppude probleemi lahendaks andmete lemmatiseerimine kasutades näiteks Texta Multilingual Processorit (MLP) või estNLTK teeki. Samuti saaks rakendada sarnase sisuga otsisõnade automaatset soovitamist.

1.1.5. Taotlustele sobivate hindajate leidmine

Probleem: Kuidas leida rahastustaotlustele selle teaduslikku sisu kõige paremini tundvad hindajad. (Punktis 1.1. nimetatud eksperkomisjonid koosnevad küll valdkonna üld ekspertidest, ent nende hulgas ei pruugi olla konkreetse taotluse teema süvaeksperte. Sellisel juhul tuleks kaastakse lisaeksperte rahvusvahelisest teadlaskonnast.)

Tegevused: Analüüsi sisendiks võeti ETISE taotluse annotatsioon ning otsiti sellele temaatiliselt kõige sarnasemaid vasteid OpenAlex andmebaasi publikatsioonide annotatsioonide seast. Suurima sarnasusega publikatsioonide autorid võiks olla sobivaks hindajaks.

Tulem: Sarnasuse otsing vektorotsingu (KNN) algoritmiga oli tulemuslik, vajadusel võib tulemust veelgi parandada märksõnaotsingu rakendamise abil. Samas tuleb arvestada, et publikatsioonide autorite seast sobivate hindajate väljasõelumine eeldab olulises mahus lisategevusi (autori täisnime tuvastamine, ei tohi olla huvide konflikti, olemas on piisav hindamise kogemus, tunneb Euroopa teadusmaastikku) ning kõik need tegevused ei ole automatiseeritavad.

Tabel 1. Ülesannete nimekiri ning analüüside lõpptulemus

Ülesanne	Analüüsi lõpptulemus
ETAG tööprotsesside ülevaade	Viidi läbi intervjuud ETAGi töötajatega, kaardistati ja kirjeldati olulisemad teemad. Analüüs on kirjeldatud peatükis 2.
Sisend- ja valdkondlike andmete analüüs	Sisendandmed hõivati ja töödeldi, saadud andmete põhjal koostati analüüsid (peatükk 3 ja peatükk 4), kirjeldati kogunenud andmete koosseis ning omavaheline vastavus (nii ETISE ja OpenAlexi andmete/valdkondade kui publikatsioonide ja valdkondade vastavus). Andmeid on piisavalt palju edasiseks analüüsiks, samuti on nende kvaliteet rahuldav.
Taotlustele valdkondade määramise analüüs	Taotluste valdkondi määrati kahe erineva masinõppe mudeli tüübi kaudu ning samuti LLM mudelite kaudu. Eesti keele mudelitega saadi üsna head tulemused - 81% taotlusi määrati õigesse valdkonda. Inglise keeles aga olid valdkondlikud andmed ebausaldusväärsed, mistõttu jäid treenitud mudelite tulemused halvemaks, isegi lisaandmete lisamisel. Ülesanne näib teostatav.
Arengukavade vastavusanalüüsi abilise analüüs	Antud ülesande puhul eksperimenteeriti kahe erineva sarnasusotsingu liigiga. Vasteid otsiti taotlustes, annotatsioonide ning sünteetiliste näidete abil. Leiti, et vektorotsingud sobivad antud otstarbeks väga hästi ja eriti juhtudel, kus arengukava sisu sõnastust või lauseehitust pole radikaalselt muudetud. Ülesanne näib teostatav.
Publikatsioonist populaarteadusliku kokkuvõtte tegemise analüüs	Publikatsioonidest eestikeelsete populaarteaduslike kokkuvõtete automaatselt tegemiseks analüüsiti viie suure generatiivse keelemudeli genereeritud kokkuvõtteid. Enamik genereeritud sisukokkuvõtteid olid lihtsamad kui algsed sisendid, kuid väikeste õigekeelsusvigadega või esines mõningaid faktivigu. Ülesanne näib teostatav.
ETIS tekstiotsingute täiendamise analüüs	Tekstiotsingute täiendamiseks pakutakse kolm varianti: sarnasusotsingute rakendamine; pealkirja ning annotatsiooni otsiväljade lemmatiseerimine; kontekstuaalselt sarnaste otsisõnade automaatne soovitamine. Ülesanne on teostatav.
Taotlustele sobivate hindajate leidmise analüüs	Sooritati ühe valdkonna seast sarnasemate publikatsioonide leidmise katse, mille käigus leiti ühele projekti annotatsioonile sarnasusotsingu kaudu viis kõige sarnasemat vastet. Ülesanne näib teostatav.
Huvide konfliktiga hindaja välistamine	Analüüsi ei sooritatud, kuna eeldused (valdkonna määramine, hindajate leidmine) on puudu.

2. Teadusagentuuri tööprotsesside ülevaade

Selles peatükis antakse põhjalik ülevaate Eesti Teadusagentuuri (ETAG) töövoogude hetkeseisust, tuginedes intervjuudele, mis viidi läbi agentuuri töötajate ja juhtidega. Analüüs käsitleb nii protseduurilisi aspekte kui ka praeguseid tööriistu, mis toetavad ETAGi igapäevast toimimist. Eesmärk on esile tuua peamised intervjuudes kõlanud tähelepanekud, tuvastada töövoogude ebatõhusust tingivad kitsaskohad, ning uurida võimalusi, kuidas AI tööriistad saaksid aidata neid probleeme lahendada.

Valdkondade määramine. Eesti Teadusagentuuri (ETAG) töövoogude üks oluline komponent on teadusrahastuse taotluste ja publikatsioonide Frascati klassifikaatoril põhinev teadusvaldkondade määramine, kuna see mõjutab otseselt kogu hindamisprotsessi ja rahastamisotsuseid. Valdkonna täpne määramine on vajalik, sest see suunab taotluse sobivasse hindamisvooru ja tagab, et õiged eksperdid seda hindaks. Praegu aga määrab valdkonna taotleja ise, mis võib tekitada mitmeid probleeme.

Peamine probleem, mida käsitleti nii intervjuudes kui ka hankes, tuleneb sellest, et taotlejad võivad määrata oma taotlusele valdkonna, millel on väiksem konkurss, kuid mis taotluse sisuga kõige täpsemini ei kattu. Nähakse, et see võib viia ebaõiglaste rahastamisotsusteni, sest hindajaks osutub sellistel juhtudel vale valdkonna ekspert.

Automaatne lahendus, mis taotluse sisu põhjal valdkondi määrab, oleks seega suureks abiks. Samas tähelepanu tuleb pöörata ka faktile, et ETAGi jaoks on tähtis, et kõikide AI lahenduste ja väljundite puhul teeks lõpliku otsuse kasutaja ise. Seega, kuigi pakutav lahendus ei takistaks taotlejat oma valdkonda AI lahenduse väljundist erinevaks muutmast, oleks mudeli pakutav valdkonnamäärang hindajatele kasulik tööriist, mis pakuks täiendavat kinnitust, et taotluse valdkond vastab tegelikkusele ja on sobivaim hindamisvooru jaoks.

Teiseks ilmneb üksikjuhtumeid, kus teatud valdkondade eksperdid väidavad, et mõne teise asutuse taotlused kuuluvad teise valdkonda. Sellega üritatakse vähendada taotleja rahastamisvõimalust ning anda enda ringkonnale seeläbi eelise. Intervjuude käigus tõstatati idee, et automaatne valdkondade määramise tööriist võiks pakkuda taotlejatele täiendavat tuge, mis annab lisatõendi, et nende taotluse valdkond on tõepoolest seotud esitatud teemaga. See suurendaks taotluste õiglast hindamist ja tagaks, et kõiki käsitletaks võrdsetel alustel.

Kolmandaks probleemiks on see, et publikatsioonide puhul jääb valdkond sageli määramata. See mõjutab eriti tugevalt analüütikute tööd, kuna valdkondade määramata jätmine raskendab teadustööde täpseks kategoriseerimiseks vajalike andmete kogumist ja analüüsi. See omakorda takistab tõhusa teaduspoliitika ja rahastamisotsuste tegemist.

Hindajate leidmine. Teine oluline töövoog ETAGis on taotlustele hindajate leidmine, mis toimub praegu manuaalselt ja on ajamahukas. Hindaja on ekspert samas valdkonnas, kellel ei tohi tekkida huvide konflikti taotluse autoritega. Tavaliselt otsitakse iga taotluse jaoks vähemalt kaks hindajat, kuid kui sobivaid kandidaate on rohkem, valitakse nad koostatud nimekirjast järjest, ilma eelisjärjekorda seadmata.

Esmalt minnakse otsinguga ETISE andmebaasi, et tuvastada, kes on varasemalt hinnanud milliseid taotlusi ja millise ekspertiisiga. Valdavalt alustatakse koostööpartneritest, kellega on varem koostööd tehtud. Kui sobivaid hindajaid ETISE kaudu ei leita, muutub protsess keerulisemaks, nõudes otsinguid ülikoolide veebilehtedelt või isiklike kontaktide kaudu. Kontaktandmete puudumine või nimekaimude eristamine võib otsingu veelgi aeganõudvamaks muuta.

Erilist tähelepanu tuleb pöörata huvide konflikti vältimisele. Näiteks tuleb välistada potentsiaalsed hindajad, kellel on olnud taotlejaga ühist teadustegevust viimase viie aasta jooksul. Oluline on ka vaadata retsensentide varasemaid hindamisi, sest eelnev koostöö ei tähenda alati, et see oli sujuv: näiteks ühelauseliste retsensioonide kirjutajad välistatakse tavaliselt edasistest hindamistest.

Seega esmane hindajate sõelumise protsess, mis võib võtta nädalaid ja päevas mitu tundi tööd, võib oluliselt kiiremaks muutuda, kui rakendada automaatne lahendus. Süsteem, mis põhineks olemasolevate andmebaaside pärimisel ja valdkonna määramise automatiseerimisel, suudaks identifitseerida potentsiaalsed eksperdikandidaadid, ning seejuures automaatselt välistada need, kelle puhul esinevad publikatsiooni kaasautorlusest tulenevad huvide konfliktid. Kuigi selline mudel ei suudaks tuvastada muid koostöövorme ja huvide konflikte peale ühise autorluse teaduspublikatsioonides, võiks see oluliselt kiirendada esialgset sõelumisprotsessi.

ETISE tekstiotsingute kasutamine ja täiendamine. ETISE tekstiotsingute praegune kasutusviis ei ole paindlik, mis piirab võimalusi leida vajalikku informatsiooni efektiivselt. Tekstiotsingu täiustamine lemmatiseerimise ja sünonüümide lisamisega – või suurte keelemudelite abil sarnasusotsingu implementeerimisega – oleks kasulik:

- analüütikutele, kes otsivad teaduslikke andmeid ja trende,
- rahastusosakonnale hindajate leidmiseks,
- kommunikatsiooniosakonnale, et leida teaduse populariseerimiseks põnevaid teadustöid,
- ministriumitele poliitika kujundamiseks ja otsuste tegemiseks

Täiustatud otsingusüsteem suurendaks eelnimetatud töö efektiivsust ja vähendaks vigu ning toetaks teadustööde hindamist, analüüsi ja teaduskommunikatsiooni.

Publikatsioonide populaarteaduslikud kokkuvõtted. Teadustöö tulemuste tutvustamine laiemale avalikkusele on teadlastele sageli tüütu lisatöö. Automaatne tööriist, mis koostab publikatsioonidest populaarteaduslikke kokkuvõtteid, aitaks teadlastel keskenduda oma põhitööle, samaaegselt tagades teadustöö parema nähtavuse ja selguse ka mitte-ekspertidele. Kuigi ETAGi töövood sellest tööriistast otseselt ei muutuks, oleks see abiks nii teadlastele kui ka teistele osapooltele, kellele töö huvi võiks pakkuda.

Vastavusanalüüsi abiline. Taotluste menetlemisel on tihti vaja hinnata nende vastavust erinevatele riiklikele arengukavadele ja strateegiadokumentidele. Praegu tehtav manuaalne vastavuse kontrollimine on ajamahukas ja keeruline, eriti kui dokumentide hulk on suur.

Vastavusanalüüsi abiline, mis suudab automaatselt leida ja esile tuua taotluste ja arengukavade vahelisi seoseid, aitaks kindlaks teha, millised teadusprojektid toetavad riiklike eesmärgi ja strateegiaid, võimaldades teha teaduspõhiseid ja strateegiliselt olulisi rahastamisotsuseid. Lisaks oleks see kasulik teadlastele, kes saaksid näidata, kuidas nende taotlus seostub kindlate arengukavade eesmärkidega.

Koosolekute automaatsed protokollid. Kuigi hankes ei ole mainitud koosolekute automaatsete protokollide loomist, ilmnes huvi sellise tööriista vastu peaaegu kõikides intervjuudes ETAGi töötajatega.

Teostatavuse hindamise analüüsi ei ole sellise tööriista puhul otseselt tarvis, sest Eestis on juba arendatud programme, mis helisisendi põhjal võimaldavad automaatset kõnetuvastust ja transkribeerimist – näiteks TTÜ poolt arendatud tekstiks.ee¹ teenus. Viimase kasutuselevõtt aitaks ETAGi töötajatel oma aega efektiivsemalt kasutada, keskenduda rohkem põhitööle/prioriteetsetele ülesannetele.

¹ <https://tekstiks.ee/> - TTÜ kõnetehnoloogia labor, rakendus eestikeelsete helisalvestiste automaatseks transkribeerimiseks

3. Sisendandmete analüüs

Antud peatükis analüüsitakse sisendandmeid, et hinnata nende usaldusväärsust ning sobivust projektis kirjeldatud lahenduste loomiseks. Täpsemalt vaadatakse andmete pikkust, mahtu, kvaliteeti ning võrreldakse OpenAlexi ja ETISE andmete omavahelist klappivust, et leida, kas antud andmed võiksid olla samaväärse kvaliteediga. Allpool on ülevaade kasutatud andmeallikate liikidest:

- annotatsioon** Lühikokkuvõtte teadusliku publikatsiooni, taotluse või projekti sisust. Analüüsis on annotatsioonid kasutatud koos publikatsiooni/taotluse/projekti pealkirjaga.
- arengukava** Strateegiliste eesmärkide kirjeldus. Analüüsis on kasutatud erinevate ministeeriumite ning riigiasutuste arengukavade tekste.
- projekti kirje** ETISes kirjeldatud teadusliku projekti kirje. Analüüsis on kasutatud CERCS klassifikaatori koodi ja nimetust, projekti pealkirja ning annotatsiooni.
- publikatsioon** Teaduslik dokument, näiteks artikkel, raamat, konverentsi ettekande kirjeldus vms. Sisendandmete analüüsis on kasutatud erinevaid publikatsiooni välju, näiteks autorite loetelu, DOI, bibliograafilised andmed, avatud ligipääs, keele nimetus/kood, ilmumisaasta, tüüp, pealkiri, OpenAlex valdkond, annotatsioon. Viimast kahte välja on kasutatud ka edasistes analüüsides.
- taotlus** Taotlus, mis kirjeldab teaduslikku projekti, mille eesmärgiks on toetuse saamine. Analüüsis on kasutatud taotluse teadusliku tausta, projekti eesmärkide kirjeldust, pealkirja, CERCS ja OpenAlex valdkonna koodi-kirjeldust ning annotatsioonide välju.

3.1 Sisendandmete kättesaamine ja töötlemine

Sisendandmetena kasutati projektide kirjeid, arengukavasid, publikatsioone ning ETISE andmete DOI põhjal kraabitud OpenAlexi andmeid. Arengukavad parsiti PDF-ist tekstilisele kujule, samuti rikastati arengukavade andmestik metainfoga (andmeallika/ministeeriumi kohta).

ETISest võeti kasutusele ainult teaduspublikatsioonide kirjed (ETISE publikatsiooni klassifikaatori järgi 1.1, 1.2, 1.3, 2.1, 2.2, 2.3, 2.4, 2.5, 3.1, 3.2, 3.3, 3.4, 3.5, 5.1, 5.2, 6.2). Lisaks jäeti ETISest ja OpenAlexist alles ainult unikaalsed kirjed, mis omakorda said olla ainult inglise või eesti keeles. Publikatsioonide hulgast eemaldati kirjete andmeväljad, mis sisaldasid ebavajalikku infot või polnud võrreldavad. Ebaolulisteks andmeväljadeks oli näiteks ETISE andmetes konverentsi kirjeldus, lõputöö tüüp jms. Lisaks vähendati erinevaid dubleeritud välju, näiteks kirjes kasutatud keeli kirjeldavatest väljadest *keeled*, *keelekood* ning *keelekood inglise keeles* jäeti alles vaid viimane.

Autorite andmed parsiti nii ETISes kui OpenAlexis lihtsaks nimedejadaks. OpenAlexis parsiti sisukokkuvõtted tekstilisele kujule. Eemaldati OpenAlexi kirjed, mis olid hiljem eemaldatud (*is_redacted*) või paratekst (*is_paratext*) nagu esilehed või sisukorrad. Eemaldati OpenAlexi kirjed,

millel pole sisukokkuvõtet või mille sisukokkuvõte oli sisuta või koosnes kirjavahemärkidest. Näiteks: *An abstract is not available for this content so a preview has been provided. Please use the Get access link above for information on how to access this content.* Samuti eemaldati pärast andmete esmast analüüsi hästi lühikeste sisukokkuvõtetega kirjed (vähem kui seitse vähemalt kolmetähelist või -numbrilist sõnet).

Juba varem DOI kaudu seotud publikatsioonide kirjetele lisati seotud projektide andmed GUID abil, sealhulgas eemaldati projektkirjetest ebavajalik informatsioon. Kolme andmestiku põhjal moodustati üks suur ühendatud andmestik (Informatsioon erinevate sisendandmete andmestike kohta **tabelis 2**).

Tabel 2. Sisendandmete tüüp, arv, allikas ja töötlus.

Allikas	Andmete tüüp	Dokumentide arv pärast töötlust	Töötlus
OpenAlex	Publikatsioonide kirjed	108 395	Andmete kraapimine DOI alusel, parsimine, filtreerimine, sisukokkuvõttete lemmatiseerimine ja tokeniseerimine
ETIS	Publikatsioonide kirjed	47 362	Andmete kraapimine, filtreerimine, parsimine
ETIS	Projektide kirjed	12 884	Filtreerimine, parsimine
Erinevad	Arengukavad	34	PDF-ide parsimine, metaandmete lisamine
Erinevad	Täispikkuses publikatsioonid automaatsete kokkuvõtete jaoks	7	PDF-ide parsimine, metaandmete lisamine

3.2 Sisendandmete kvaliteedi hindamine

Sisendandmete kvaliteeti hinnati nii automaatselt kui käsitsi. Kuna valdkonnaandmeid ei ole võimalik otseselt kõrvutada, siis otsiti välja erinevad metaandmete kategooriad, millel oli mõlemas andmestikus olemas võrreldavad väärtused. Kõige olulisemaks kategooriaks oli publikatsioonide autorite info, kuna autorite nimedega seonduvad järgnevad ülesanded hindajate leidmiseks ning konfliktide vältimiseks.

Võrreldavad publikatsiooni kirje metaandmed olid:

- publikatsiooni autorid,
- publikatsiooni pealkiri,
- ilmumisaasta,
- bibliograafilised andmed nagu väljaande number, algus- ning lõpulehekülg ja ISSN,
- kas allikas oli avatud juurdepääsuga,
- publikatsioonis kasutatud keel(ed).

Antud metaandmete kategooriate kaudu hinnati kattuvuse määra, kus igale kirjele arvutati keskmine kattuvuse määr. Samuti uuriti, kas ETISes määratud publikatsiooni tüüp kattub OpenAlexi andmetüübiga, kuid seda näitajat kattuvuse arvutusse ei kaasatud, kuna polnud teada, kas andmed on võrreldavad või mitte.

3.2.1 Sisendandmete hindamismeetodid

Hinnati, kas kaks metainfo osa kattuvad täielikult või osaliselt. Täieliku kattuvuse puhul tagastas kirjeid võrdlev skript antud kategoorias 100 (tähtsusega 100% kattuvus). Eri metaandmete tulemuste põhjal leiti iga dokumendi jaoks metaandmete keskmine kattuvus protsendi kujul. Puuduv info sai automaatselt väärtuse 0. Tekstiväärtuse osalise kattumise puhul arvutati erinevus ETISE kirjest Levenshteini kauguse abil (mitu muudatust on vaja ühele sõnele/arvule teha, et saada teine sõne/arv), mida kasutati kattuvuse protsentarvutuses. Järgneb näide Levenshteini kauguse kaudu kattuvuse arvutamisest:

ETIS pealkiri – *Õpikuteksti analüüs kirjaoskuse omandamise kontekstis*

OpenAlex pealkiri – *Opiteksti Analuus Kirjaoskuse Omandamise Kontekstis*

Levenshteini kaugus: 9

Kattuvuse protsent: 83.02%

Järgnevad kirjeldused täpsematest automaatsetest hindamismeetoditest andmeliigiti:

Autorid

Autorite puhul võeti arvesse ETISE kasutajatega seotud autoreid, seega otsiti, kas väike alamhulk autoreid on OpenAlexi autorite nimekirjas olemas. Kuna töö eesmärgiks seatud rakenduste eelduseks on leida põhiliselt Eestiga seotud teadlasi, siis eeldati, et ETISE kasutajatega seotud päringud on eelistatud. Nimekujud puhastati kirjavahemärkidest. Mõnes kirjes esines sama nimi korduvalt, seetõttu otsiti unikaalseid nimesid. Otsiti ka osaliselt kattuvaid nimesid regulaaravaldiste abil.

Pealkiri

Pealkirja puhul arvutati Levenshteini kaugus erinevate pealkirjavormide vahel, mille alusel arvutati kattuvus.

Ilmumisaasta

Ilmumisaasta puhul tehti lihtne protsentarvutus kasutades kahe aastaarvu vahet.

Bibliograafilised andmed

Bibliograafilistes andmetes parsiti ISSN andmeid, kuna osa publikatsioone oli mitme/täpsema ISSN numberiga. Vaadati, kas saadud numbrid kattuvad mingil määral OpenAlexi andmetega. Algsu- ning lõpuleheküljenumbrite puhul kasutati protsentarvutust vahega, samuti väljaande numbri puhul.

Juurdepääs

Avatud juurdepääs oli mõlemas allikas kas tõene või väär (tõene, st publikatsioon on avatud juurdepääsuga), seega sai tulemus olla ainult 100 või 0.

Keeled

Keelte puhul teisendati ETISE andmed keelekoodideks, kuna seal oli koodidega ühes ka keele nimetused (näiteks "inglise, ET"). Kui üks keeltest klappis, siis loeti andmed kattuvaks.

Tüüp

Võrreldi ETISE publikatsiooni tüüpi OpenAlexi publikatsiooni tüübiga.

Kuna publikatsiooni tüübid polnud otseselt seotud, leiti tekstitüübid, mis sagedamini kattuvad.

3.3 Sisendandmete kvaliteedi hindamise tulemus

3.3.1 Sisendandmete kvaliteedi hindamise üldtulemused

Saadud tulemustest tekitati tabel: Tabel üldtulemustega². Tulemuste põhjal teostati veel puhastustöid, seega kasutatud sisendandmete arv muutus. Tabelist nähtub, et vähemalt keskmise kattuvuse järgi üle 79% kirjetest kattuvad suurel määral. Vähem kattuvaid dokumente on alla 21%.

Ilmumisaasta kattus enamikel juhtudel üsna hästi, vaid kümme juhtu oli vale ilmumisaastaga. Esines kümme kirjet, millel ETISes ja OpenAlexis oli erinev ilmumisaasta arv (**tabelis 3**). Kahel kõige vähem kattaval aastaarvul oli ilmselt aastaarv vale lisandunud täiesti suvalise ISSN numbri tõttu, ülejäänutel ei suuda tuvastada põhjust erinevuse tekkimiseks.

Tabel 3. Erinevate ilmumisaastate loend kattuvusprotsendiga.

Ilmumisaasta ETISE järgi	Ilmumisaasta OpenAlexi järgi	Aastaarvude kattuvusprotsent (%)
2011	1970	98
2017	1969	98
2016	2003	99

² ETIS ja OpenAlex metaandmete automaatse võrdluse tulemused.xlsx

2004	2017	99
2014	2001	99
2003	2019	99
2011	1999	99
1991	2002	99
2010	2021	99
2010	2021	99

Samuti kattus väga hästi **keelte** esinemine publikatsioonis. Erinevused põhjustas see, et OpenAlexil oli automaatselt määratud vaid üks keel sisukokkuvõtte ja/või pealkirja põhjal. Samas võib sisukokkuvõtte olla mitmes keeles. Valdav enamik publikatsioone kogunenud andmestikus on ingliskeelsed.

Pealkirjade puhul ei olnud väga suuri eksimusi palju. Erinevused tekkisid näiteks alampealkirja esinemisel kas ETISes või OpenAlexis, samuti oli OpenAlexi pealkirjades HTML koodi või oli eestikeelse artikli pealkiri OpenAlexis kas inglise keeles või nii eesti kui inglise keeles. Ka pealkirjades esineb vigu, kus DOI kaudu seostatud kirjed pole tegelikult seoses, vaid on lihtsalt samas väljaandes esinenud.

Avatud juurdepääsuga publikatsioonide puhul eksis OpenAlex pigem nende kirjete puhul, mis olid kättesaadavad ETISe järgi.

Kõige vähem kattuvad andmed olid **bibliograafilised metaandmed** ning autorite nimed. Osa kirjetel polnud üldse bibliograafilisi metaandmeid küljes, mõnel seetõttu, et tegu oli pre-prindiga. Algu- ja lõppleheküljed võisid olla suvalised, näiteks ühel juhul oli nii algus- kui lõpplehekülg *ddw134*. Võrdlemisprobleem tekkis raamatukirjetega, kuna OpenAlexis ISBN välja ei esine. Iseenesest ei tohiks bibliograafiliste andmete puudumine järgnevaid tööülesandeid mõjutada, kuna järgnevates ülesannetes on pigem vaja sisukokkuvõtteid ning autorite nimesid, küll aga näitab see üldist andmete kvaliteeti.

3.3.2 Publikatsioonide tüüpide hindamine

Kuna publikatsioonide tüübid otseselt ei kattunud, siis koostati selle jaoks võrdlustabel: Publikatsioonide tüüpide võrdlustabel³. OpenAlex saab tekstitüübi osas andmeid [Crossref-ilt](#).

Publikatsioonide tüüpide võrdlustabelist ilmneb, et enamik kogunenud andmetest on teadusartiklid, selle publikatsioonitüübi puhul oli suures osas ETISe ja OpenAlexi kirjetes üksmeel. Teine tüüp, konverentsiesitlus, sobis mingil määral kokku OpenAlexi tekstitüübiga konverents-artikkel (i.k. *proceedings-article*). Ka artiklid/peatükid raamatutes sobisid OpenAlexi

³ ETIS ja OpenAlex publikatsiooni tüüp.xlsx

sarnase tekstitüübiga raamat-peatükk (i.k *book-chapter*). Ülejäänud publikatsioonitüüpide osas polnud väga suurt kattuvust. Otseselt pole see ka andmete kvaliteedi osas oluline, aga analüüsisist nähtub, et publikatsioonide tüübid on väga sarnased.

3.3.3 Autorite nimede hindamine

Autorite nimede puhul selgus, et mitmes suures geeniuringus on Eesti teadlased osalenud väiksema uurimisgrupi raames. OpenAlexi kirjetes aga uurimisgrupp ega selle uurimisgrupi liikmed autoritena ei figureeri, seega on üsna mitu kirjet, kus Tõnu Esko, Andres Metspalu jt on ETISes autorid, kuid OpenAlexis mitte. Seetõttu tekkis palju kirjeid, kus ETISe autoreid ei leitud üldse OpenAlexi kirjetest.

Samuti mõjutas autorite nimede leidmist OpenAlexi nimevormide paljusus – kuna andmed on automaatselt kogutud, pole autori nime osas selgelt vormi. Seeõttu oli palju osalisi klappimisi, kus autor tegelikult leiti näiteks initsiaalide ja perenime kaudu või ilma keskmise nimeta. See tähendab, et Levenshteini kaugus võib näidata samas kirjes leitud nimede vahelisi erinevusi suuremana kui need tegelikult on. Osalisi klappivusi automaatselt samaks isikuks määrata on aga ohtlik: näiteks kui publikatsioonil on kaks autorit, kes on mõlemad sama perenimega ja sarnaste eesnimedega, siis on ka nimede klappivus omavahel väga suur ning initsiaali järgi on võimatu määrata, kumba mõeldakse.

Vanemate kirjete puhul oli ka probleemiks translitereerimine, näiteks Enn Tõugu puhul on OpenAlexis autorinime vormiks *Enn Tyugu* või *E.H. Tyugu*. Esines ka vähemalt üks juht, kus autori nimi oli valesti kirjutatud või esines nimest teine vorm. Kirjed, milles autori nimed üldse ei klapi, võivad olla vigased väljaande ning spetsiifilise artikli DOI kattumise tõttu.

3.3.4 Sisukokkuvõtete hindamine

OpenAlexi sisukokkuvõtete kvaliteet on oluline taotluste valdkonna tuvastamise prototüübi loomisel, seega sisukokkuvõtteid vaadati kvaliteedi hindamiseks käsitsi. Kokku vaadati läbi üle tuhande sisukokkuvõtte.

Enamik läbivaadatud sisukokkuvõtetest olid kvaliteetsed ning sobiksid kasutada näiteks valdkonna määramiseks. Väiksem osa sisukokkuvõtetest olid sisutud, sisaldades näiteks viidet antud publikatsioonile, publikatsiooni avaldamise informatsiooni, autori nime või arusaamatuid tähemärkide jadasid (enne keeletuvastust). Näited sisututest sisukokkuvõtetest:

"Text, Context and Performance: Gagauz Folk Religion in Discourse and Practice." Folklore, 124(3), pp. 369–370

–

Saabunud / Received 07.11.2019 ; Aktsepteeritud / Accepted 19.11.2019 ; Avaldatud veebis / Published online 29.11.2019 ; Vastutav autor / Corresponding author: <e-mail>

–

0112ÿ 4ÿ 56 7 8 ÿ 199 ÿ 7 9 ÿ 99ÿ ÿ 9ÿ ÿ 2 7 ÿ 9ÿ ÿ 2 7 ÿ !"#\$ %&'()* \$ÿ + "(,ÿ ; -# .ÿ /01 1ÿ)1 2"3ÿ \$
 ",ÿ)%(4 #4 "ÿ % #ÿ \$ #""ÿ)5%-)"ÿ %\$ ÿ)%(4 #4 "ÿ (,ÿ 678 9: ÿ <=>?=>@ ÿ A: ÿ <BCD @ ÿ

Veel oli üks väiksem osa sisukokkuvõtteid, kus oli kas lõpetamata sisukokkuvõtte algus või mõni muu hästi väike osa sisukokkuvõttest. Osaliselt kaduma läinud sisukokkuvõtted:

Ülevaade. Artiklis vaatlen füüsilise teatri arengut maailma

–

Tänapäeva teaduses on

–

146) või veel kuidagi teisiti

Samuti leiti ka üks sisukokkuvõte, mis sisaldas mitte-sõnalisi sümboleid. Näide puhastamata/koodiga sisukokkuvõttest:

   Herewith, we continue to upgrade Estonian checklist of lichenized, lichenicolous and allied fungi and report twelve fungal species as new for Estonia, of them eleven are lichenized and one is lichenicolous fungus.  

ETISest leitud sisukokkuvõtted olid samaväärse kvaliteediga kui OpenAlexi sisukokkuvõtted, kuid neid oli hästi vähe, 27 tükki eesti keeles ja 177 inglise keeles. Seetõttu neid edasisesse analüüsisse ei kaasatud.

3.4 Hinnang sisendandmete kvaliteedile

Kokkuvõtteks võib öelda, et andmete kvaliteet on hea (küll teatud määradega). Näiteks autorite nimede puhul tuleb väga täpselt läbi mõelda, kuidas asendada puuduvaid andmed (nt need autorid, kes mõnes väiksemas uurimisgrupis publikatsioonis kaasa teevad) ning kuidas andmeid töödelda, et saada kõik nimekujud kätte.

Parimate tulemuste saamiseks tuleks ette võtta täpsem sisukokkuvõtete puhastamistöö. Osa sisukokkuvõtete puhastusest sooritati pärast analüüsi:

- Osaliselt kaduma läinud sisukokkuvõtted filtreeriti välja pikkuse alusel, kehtestades sisukokkuvõttele miinimumpikkuse – sisukokkuvõte peab sisaldama rohkem kui seitse sõnet ja vähemalt ühe sõne pikkus peab olema pikem kui kolm tähemärki.
- Mitte-sõnalistest sümbolitest sisukokkuvõtted on välja võetud keeletuvastuse abil.
- Samuti eemaldati kirjed sisukokkuvõtete, milles oli vaid publikatsiooni avaldamise kirjeldus.

Osa keerulisemaid puhastustöid, mis jäid tegemata:

- Ilmselt saaks regulaaravaldiste kaudu kätte muid metaandmeid sisaldavad sisukokkuvõtted.

- Koodi või sümboleid sisaldavaid kirjeid saaks puhastada samuti regulaaravaldistega, kuid seda tuleks ettevaatlikult, et ei läheks kaduma olulist informatsiooni. Näiteks matemaatilised sümbolid, mis on HTMLiga väljendatud, võivad olla sisukokkuvõttes tähenduslikud.

4. Valdcondlike andmete analüüs

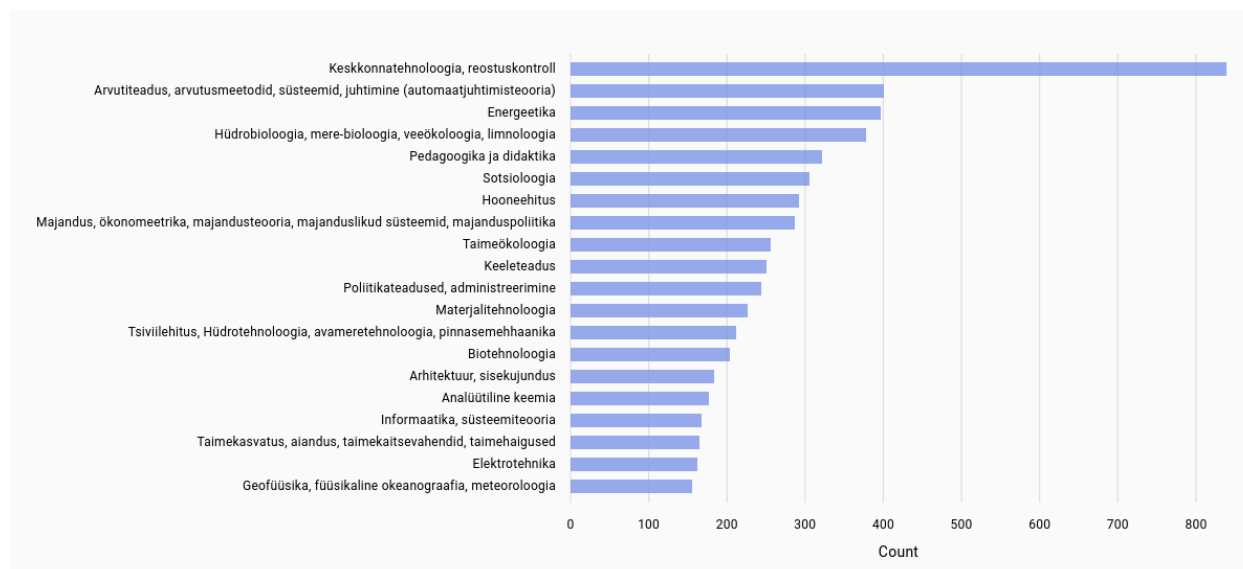
Analüüsis hinnati sisendandmetes sisalduvate valdkondlike andmete kvaliteeti, muuhulgas publikatsioonile määratud valdkonna vastavust publikatsiooni annotatsioonile. Uuriti eri valdkondade hierarhiaid ning võrreldi omavahel ETISes kasutatava *Common European Research Classification Scheme* ja OpenAlexis kasutatavate Scopus klassifikaatorite valdkondi, samuti vaadeldi märksõnu. Valdcondi võrreldi nii automaatselt kui käsitsi, et teada saada, kas valdkonnamääratlused on piisavalt kvaliteetsed ning piisavalt suure mahuga, et nende põhjal masinõppe mudeleid treenida.

4.1 Valdcondlike andmete kvaliteedi hindamine

Valdkondlike andmete kvaliteeti hinnati nii automaatselt kui käsitsi. Automaatselt võrreldi OpenAlexi *secondary topic subfield* tasemel olevaid valdkonnamääratlusi (edaspidi OpenAlexi alamvaldkonnad või valdkonnad), ETISe publikatsioonide märksõnu ning projektide *Common European Research Classification Scheme* valdkondi (edaspidi CERCS valdkondi). Samuti vaadati valdkondlike andmete esinemissageduse jaotust OpenAlexi alamvaldkondades, kus kasutati ka kõige tõenäolisemat valdkonda (*primary topic subfield*).

Kokku oli projektide automaatses võrdluses 8926 kirjet. Kirjete arv tulenes projektidest, mis seostusid otseselt mõne publikatsiooni kirjega ning olid 100% seoses ühe valdkonnaga. Igaks juhuks tuleb toonitada, et projekti valdkonnal ei pruugi olla projektiga seotud publikatsiooni annotatsiooniga üksühest seost. Projektidel võis olla 1–8 CERCS valdkonda, OpenAlexi alamvaldkondi oli üks kuni kolm. Ühel kirjel võis olla mitu sama väärtusega OpenAlexi alamvaldkonda (nt *Genetics* ja *Genetics*), kuna OpenAlexis on kõige tähtsam *topic* valdkond, ning on loogiline, et teemat määrates võivad kaks erinevat teemat olla sama alamvaldkonna all. Samamoodi võisid olla ühe publikatsiooni kirjega seotud projektid samast valdkonnast.

Joonisel 1 on näidatud kõige sagedamini esinenud valdkonnad puhastatud projektide hulgas. Nagu näha, siis kõige sagedamini esinenud valdkonnad on keskkonnatehnoloogia ja reostuskontroll; arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria); energeetika.



Joonis 1. Kõige sagedasemate valdkondade nimetused eesti keeles puhastatud projektide hulgas.

Lisaks koostati kaks tabelit projektide valdkondade ning OpenAlexi alamvaldkondade koosinemiste kohta. Üks tabel hõlmab kõiki valdkonnamääratlusi ühe kirje kohta (lõige **tabel 4**), teine tabel hõlmab erinevaid valdkonna-projekti kombinatsioone, mis koos esinesid, mida sortiti sageduse alusel (lõige **tabel 5**). **Tabelist 5** eraldati omakorda need kombinatsioonid, kus oli sarnane tekstiline osa (näiteks *Spectroscopy* ja *THz time-domain spectroscopy*, tabel asub siin⁴).

Tabel 4. Valdkonnamääratlused ühe kirje kohta, lõige tabelist projektiga kirjeid võrdlev tabel⁵.

Täispikkuses tabel on kättesaadav lingi kaudu.

CERCS kood	ETIS projekti valdkonnad (inglise keeles)	OpenAlex alamvaldkonnad tõenäosusega
['S260']	['Psychology']	{'Clinical Psychology': '0.9995', 'Applied Psychology': '0.9967'}

Tabel 5. Valdkonnamääratlused ühe kirje kohta, lõige tabelist koos esinevad valdkonnad⁶.

Täispikkuses tabel on kättesaadav lingi kaudu.

ETIS projekti valdkond (inglise keeles)	CERCS kood	OpenAlex alamvaldkond	Esinemissagedus
Computer science, numerical analysis, systems, control	P170	Artificial Intelligence	420

⁴ Valdkonnad tekstiline kattuvus.xlsx

⁵ ETIS valdkondade ja OpenAlex alamvaldkondade koosinemine.xlsx

⁶ ETIS projekti valdkonna ja OpenAlex alamvaldkonna koosinemine.xlsx

Märksõnakirjeid oli 15 212. Kuna märksõnad olid erinevalt vormistatud, siis ei hakatud märksõnu eraldi parsima, vaid need jäid üldjuhul loeteluks. Märksõnade ja alamvaldkondade koos esinemise kohta moodustati samuti tabel (lõige **tabelis 6**).

Tabel 6. Märksõnadega koos esinevad valdkonnad ühe kirje kohta, lõige tabelist valdkonnad ja märksõnad⁷. Täispikkuses tabel on kättesaadav lingi kaudu.

ETIS märksõnad vabatekstina	ETIS märksõnad inglise keeles	ETIS märksõnad	OpenAlex alamvaldkonnad
['Biological pathway', 'Major depressive disorder', 'NETRIN1', 'Polygenic risk score', 'Thalamic radiations', 'White matter']	[]	[]	{'Radiology, Nuclear Medicine and Imaging': '0.9989', 'Genetics': '0.991'}

4.2 Valdkondlike andmete kvaliteedi hindamine

Automaatselt koostatud tabelite järgi nähtub, et potentsiaalset kattuvust on nii projektide valdkondade ja alamvaldkondade kui ka märksõnade ja alamvaldkondade vahel. Osaliselt kattuva tekstilise sisuga on 65 valdkonda ja 421 märksõna.

Mõni valdkond paistab täielikult kattuvat, kuid teiste puhul tekib probleem, kuna CERCSi ja OpenAlexi valdkondade detailsusaste on erinev. Ühe CERCS valdkonna kohta võib olla mitu OpenAlexi alamvaldkonda ja vastupidi, näiteks kattuva tekstilise sisuga tabelis on OpenAlexis "Law" (õigus) ja CERCS valdkonnas "History of Law" (õigusajalugu). Märksõnade hulgas on valdavalt ülekaalus samuti erineva detailsusastega vasted.

OpenAlexi valdkondade jaotus on väga ebaühtlane: kõige levinumad valdkonnad esinevad harvematega võrreldes kuni 500 korda sagedamini. See võib mudeldamisel põhjustada sagedamate valdkondade eelistamist olenemata sisendist. Kui vaadelda kirjete kõiki võimalikke valdkondi, siis on valdkondade esinemise ebavõrdsus veidi väiksem, kuid võib siiski tulemusi mõjutada.

⁷ ETIS märksõnade ja OpenAlex valdkondade koosinemine.xlsx

4.2.1 CERCS valdkonnad ja OpenAlexi alamvaldkonnad

4.2.1.1 Automaatne võrdlus

Projektiga kirjeid võrdleva tabeli⁸ (**tabel 4**) alusel saab näha, millised publikatsiooni kirjega seonduvad projekti valdkonnad ning sama publikatsiooni kirjega seonduvad OpenAlexi alamvaldkonnad esinevad koos. OpenAlexi alamvaldkondade juures on ära toodud ka iga alamvaldkonna tõenäosus (nullist üheni).

Analüüsist nähtub, et projektide valdkonnad on OpenAlexi detailsemate valdkondadega võrreldes laiapinnalisemad. Näiteks leidub publikatsioonides üle 120 kirje, mis seostuvad analüütilise keemia projektidega. Kuigi analüütiline keemia on ka OpenAlexi valdkondade seas olemas, siis OpenAlexi alamvaldkonnad samadel publikatsioonidel ei piirdu ainult selle valdkonnaga, esindatud on ka farmakoloogia, biomeditsiinitehnika, toiduteadus, keemiline materjaliteadus, entomoloogia, molekulaarbioloogia jms. OpenAlexi valdkondade hulgast puuduvad täielikult näiteks “folkloristika” ja “üldpatoloogia”. Seega tuleb tõdeda, et CERCS ja OpenAlexi valdkondade klassifikatsioonid ei kattu täielikult, mistõttu ei saa klassifikaatoreid automaatselt ühendada.

Võrreldi ka seoseid iga kirje võimalike projektide ning võimalike alamvaldkonna variantide vahel. Analüüsisiti, millised OpenAlexi alamvaldkonnad seostuvad projekti kaheksa kõige sagedasema valdkonnaga (vt **tabel 7**). Võrdluseks võeti koos esinevate valdkondade tabel. Vaadeldavatel valdkondadel on vähemalt paar erinevat sagedasemat vastet OpenAlexi taksonoomias (mida esineb rohkem kui 5%), palju on ka OpenAlexi alamvaldkondi, mis esinevad vaid korra (1%–8% kõigist valdkondadest olenevalt projekti valdkonnast). Projektile vastav kõige sagedasem alamvaldkond hõlmas vaid 15%–35% kõigist pakutud valdkondadest.

Tabel 7. Detailsem ülevaade sagedamini esinevatest valdkondadest.

CERCS valdkond	Kolm sagedamat OpenAlexi alamvaldkonda	OpenAlexi alamvaldkonna esinemisprotsent
Keskonnatehnoloogia, reostuskontroll	Ecology	15,8%
	Global and Planetary Change	9,9%
	Oceanography	9,9%
Arvutiteadus, arvutusmeetodid, süsteemid, juhtimine (automaatjuhtimisteooria)	Artificial Intelligence	27,9%
	Computational Theory and Mathematics	13,3%
	Information Systems	10,6%

⁸ ETIS valdkondade ja OpenAlexi alamvaldkondade koosinemine.xlsx

Energeetika	Electrical and Electronic Engineering	25,0%
	Renewable Energy, Sustainability and the Environment	14,1%
	Materials Chemistry	12,4%
Hüdrobioloogia, mere-bioloogia, veeökoloogia, limnoloogia	Oceanography	27,4%
	Ecology	18,1%
	Global and Planetary Change	14,3%
Pedagoogika ja didaktika	Education	34,0%
	Developmental and Educational Psychology	11,7%
	Social Psychology	9,3%
Sotsioloogia	Sociology and Political Science	29,9%
	Political Science and International Relations	9,6%
	Communication	7,0%
Hooneehitus	Building and Construction	34,2%
	Environmental Engineering	13,9%
	Social Psychology	5,0%
Majandus, ökonomeetrika, majandusteooria, majanduslikud süsteemid, majanduspoliitika	Economics and Econometrics	15,6%
	Strategy and Management	14,4%
	Accounting	9,3%

Eelnevast tabelist nähtub, et enamikel juhtudel on kõige sagedamini seostatud alamvaldkond kõige paremas vastavuses CERCS klassifikatsiooniga, välja arvatud arvutiteaduse puhul, kus teine variant oleks veelgi paremas vastavuses. Pole üht kindlat valdkonda, mis ühegi projekti valdkonna puhul esile kerkiks (protsendid on üldiselt madalad), kuid see võib tuleneda sellest, et arvestati kõiki võimalikke valdkondi, mitte ainult kõige tõenäolisemaid. "Keskkonnatehnoloogia" ning "hüdrobioloogia" said mõlemad täpselt samad vastused, aga võib arvata, et need valdkonnad kattuvadki rohkem võrreldes teistega.

4.2.1.2 Keelevaldkonna lühikokkuvõtete võrdlus

Uurisime käsitsi OpenAlexi valdkondi ning valdkonna seost lühikokkuvõtte sisuga, mille põhiliseks alamvaldkonnaks oli määratud **keeleteadus ja lingvistika** (*Language and Linguistics*). Keelevaldkond valiti, kuna sellega on analüüsi autoril kõige suurem varasem kogemus.

Lisaks **keeleteaduse ja lingvistika** valdkonnale on andmestikus olemas ka alamvaldkond **lingvistika ja keeleteadus** (*Linguistics and Language*), mis tuleneb OpenAlexi kasutatud SCOPUS ASJC klassifikaatorist⁹, kus on mõlemad valdkonnad olemas. See viitab, et valdkondades võib veel olla sarnaseid rööpseid kategooriaid. Seetõttu keskenduti selliste kirjade analüüsile, mille põhilises alamvaldkonnas sisaldus sõna **keel** (*Language*). 496 kirjest vaadati läbi 200. Läbivaadatud sisukokkuvõtete põhjal ei tuvastatud olulist erinevust kahe alamvaldkonna vahel, kuigi valdkonnad peaks olema sisuliselt erinevad.

Enamikul läbivaadatud lühikokkuvõtetest oli selge seos keele- või kirjandusteadusega. Mitmed keelevaldkonna lühikokkuvõtted olid aga seotud tõlkimisega, hariduse, semiootika, filosoofia või folkloristikaga, ning need publikatsioonid võiksid kuuluda mõnda teisse valdkonda. Samas on paljude valdkondade piirid hägusad ning teadustöö võib kuuluda mitmesse valdkonda korraga.

Lühikokkuvõtted olid enamjaolt kvaliteetsed. Keeles ja Kirjanduses ilmunud artiklite lühikokkuvõtted olid tihti osalised, vahel oli see ka markeeritud [...] märgiga. Näiteks: "Siinses artiklis on vaatluse all töölikirjandus ning selle mõiste määratlemisega seotud probleemid. Kas on võimalik üheselt"

Samuti oli üks kirje sattunud andmestikku vähemalt kaks korda (DOIga <http://doi.org/10.7592/methis.v17i21/22.14582>/<https://doi.org/10.7592/methis.v17i21/22.14582>), kuid see ei tohiks tulemusi oluliselt mõjutada.

4.2.2 Märksõnad ning OpenAlex alamvaldkonnad

Publikatsioonidele lisatud märksõnad on üsna teadustöö-spetsiifilised. Kuigi mingi osa hõlmab ka laiemalt valdkonna kirjeldamist, siis üldiselt on märksõnadeks üsna spetsiifilised terminid, seetõttu on keeruline neid OpenAlexi valdkondadega otseselt võrrelda.

Näiteid märksõnadest:

*16S rRNA and occupants' age | airborne microbiome | meteorological data | Northern Europe
Ageing | Chronic pain | Cognition | Epidemiology | Male health | Population survey*

—

'agricultural catchments', 'concentrations', 'losses', 'monitoring', 'phosphorus', 'temporal trends'

⁹ https://service.elsevier.com/app/answers/detail/a_id/15181/supporthub/scopus/ - Elsevier, SCOPUS klassifikaatori loend

Märksõnade analüüsist nähtub, et mingit osa märksõnu saaks valdkondade tuvastamise protsessis kasutada juhul, kui sisukokkuvõtte pole piisavalt kvaliteetne või informatiivne. Märksõnade kasutamise puhul tuleks siiski kaaluda, kuidas need samale kujule viia (kas märksõnaeraldajaks on koma, püstkriips vms) ning kontrollida, et märksõnu oleks andmetes piisavalt.

4.3 Lõpphinnang valdkondlikele andmetele

Automaatsetest ning käsitsi võrdlustest nähtub, et OpenAlexi valdkondlike andmete kvaliteet on pigem rahuldav. Otsene vastavus CERCS klassifikatsioonile puudub. Samas on andmeid üsna palju, mis on masinõppe kasutamise suhtes soosiv.

Kuna erinevatele keeltele tuleks treenida keeletespetsiifilised mudelid, siis tekib probleem eestikeelsete andmetega, mida on väga vähe (umbes 2% kogu andmestikust, st umbes 1000 kirjet). Mudelite treenimiseks kasutati eesti keele puhul projektide annotatsioone. Kui andmeid on vähe ning nende kvaliteet ebaühtlane, siis tuleks tulevikus uurida võimalusi rohkem andmeid juurde saada või andmete kvaliteeti parandada. Võimalik oleks kasutada ETISe projektide ingliskeelsete annotatsioonide andmeid ja/või masintõlget, et andmete hulka suurendada.

Peamine probleem andmete puhul on selles, et sama suuri kättesaadavaid alternatiivseid andmestikke pole palju. Alternatiivne andmestik Elsevier avatud juurdepääsuga artiklitest koos ASJC kesktaseme valdkonnaga (40 000 kirjet, sarnased väärtused kui OpenAlexi *field* tasemel) on kättesaadav Huggingface lehel¹⁰. Selle puhul on aga andmed ainult inglise keeles ning valdkonnatase on üldisem. Samuti võib leida andmeid ühest Kaggle võistlusest¹¹, kus oli 100 000 kategoriseeritud sisukokkuvõtet, kuid need andmed on ilmselt vananenud ja valdkondi on vaid neli.

Ei eksisteeri ka alternatiivseid olemasolevaid tööriistu, mis eestikeelsete tekstide valdkonna tuvastamiseks sobiksid. Varasemalt on tehtud teaduslikel tekstidel masinõppe kasutamiseks näiteks BERTeley¹², sciBERT¹³ ja teematuvastuseks on eraldi BERT mudel BERTopic¹⁴. Kõik need tööriistad eeldavad, et sisendtekstid on ainult inglise keeles, mis ei pruugi sobida ETISe taotluste töövoogu.

¹⁰ <https://huggingface.co/datasets/orieg/elsevier-oa-cc-by> - Elsevier andmebaasist kogutud andmestik HuggingFace keskkonnas

¹¹ <https://www.kaggle.com/c/abstract-classification/data> - Kaggle võistluse sisukokkuvõtete andmed

¹² Chagnon, Eric jt 2024. [Benchmarking topic models on scientific articles using BERTeley](#). Natural Language Processing Journal 2024 (6).

¹³ Beltagy, Iz; Lo, Kyle ja Cohan, Arman 2019. [SciBERT: A Pretrained Language Model for Scientific Text](#). EMNLP 2019

¹⁴ Grootendorst, Maarten 2022. [BERTopic: Neural topic modeling with a class-based TF-IDF procedure](#). arXiv:2203.05794

5. Taotlustele valdkondade määramise analüüs ja prototüüp

Analüüsi käigus uuriti, milline lähenemine võiks sobida kõige paremini taotluste valdkonna automaatseks määramiseks. Valdkonna automaatset määramist katsetati nii BERT-klassifitseerijaga, Tagger Groupi kui ka suurte generatiivsete keelemudelitega.

5.1 Alusandmed ja definitsioonid

Tabel 8. Taotlustele valdkondade määramise alusandmete keel, dokumentide arv ja kirjeldus.

Keel	Dokumentide arv	Kirjeldus
eesti	18 551	Lemmatiseeritud projektide annotatsioonid koos CERCS valdkondadega (kõigi valdkondadega, mitte ainult 100% ühe valdkonnaga projektid).
inglise	46 588	Publikatsioonide lühikokkuvõtete tokeniseeritud tekst ja OpenAlexi valdkonnad.

Eesti keele puhul kaasati kõik projektid. Alusandmed lõodi keele alusel lahku, kuna valdkonnad ja andmete žanr (kas publikatsiooni lühikokkuvõte või projekti annotatsioon) on erinevad. Samuti erineb andmete kogus, inglise keeles on rohkem kui kaks korda enam andmeid kui eesti keeles (**tabel 8**). Lisaks oli veel üheksa taotlust testimiseks (**tabel 9**), millel on nii eesti- kui ingliskeelsed annotatsioonid, ingliskeelsed teadusliku tausta kirjeldused, projekti eesmärgid ning OpenAlex ja CERCS valdkonnad.

5.2 Andmete töötlus

Kõik andmed asusid töötlemise ajal tööserveris, millele pääseb ligi kas Texta kontori võrgust või VPNi kaudu. Andmed tokeniseeriti ja lemmatiseeriti kasutades Texta Multilingual Processor (MLP)¹⁵ teeki. Taotlustest eemaldati kõigilt tekstiväljaldelt HTML-kood ning lemmatiseeriti eestikeelne annotatsioon.

Mõlemad andmestikud jaotati treenimiskogumiks ja testimiskogumiks, võttes umbes 20% testandmestikuks. OpenAlexi andmete puhul arvestati, et mudelite treenimisel on võimalik kasutada nii üht põhivaldkonda (*primary topic subfield*) kui ka kõiki ülejäänud valdkondi (*secondary topic subfield*). Projektidest tulenenud andmetel kuulus aga suurem osa annotatsioone ühte CERCS valdkonda.

¹⁵ <https://pypi.org/project/texta-mlp/> - Texta Multilingual Processor, Pythoni teek keeleandmete töötlemiseks

Valdkondade märksõnade puhul kasutati eestikeelsete andmete jaoks CERCS koodi ning ingliskeelset valdkonna nimetust, samas kui inglise keeles kasutati ainult OpenAlexi valdkonna nimetust. Andmestike kirjeldused on esitatud järgnevas tabelis:

Tabel 9. Dokumentide jaotus valdkonna tüübi põhjal treenimis- ning testandmestikku.

Testandmestiku nimetus	Alusandmed	Valdkonna tüüp	Dokumentide kogus treenimis-andmestikus	Dokumentide kogus testandmestikus
<i>OA_primary</i>	OpenAlex	Põhivaldkonnad	37 178	9295
<i>OA_secondary</i>	OpenAlex	Kõik valdkonnad	37 270	9318
<i>P_secondary</i>	Projektid	Kõik valdkonnad	14 840	3711
<i>Application</i>	Taotlused	OpenAlex/CERCS valdkonnad	-	9
<i>Parasite_primary</i>	OpenAlex	Põhivaldkonnad	19 354	71
<i>Parasite_secondary</i>	OpenAlex	Kõik valdkonnad	19 354	71
<i>Psych_new</i>	OpenAlex	Põhivaldkonnad/ Kõik valdkonnad	35 982	5697
<i>Psych_old</i>	OpenAlex	Põhivaldkonnad/ Kõik valdkonnad	2235	5697

5.3 Meetodite võrdlus

Klassifitseerimiseks treeniti kahte tüüpi mudeleid: Transformeri-põhiseid BERT mudeleid ning statistilisi Tagger Group mudeleid. Mudelite tulemused valideeriti testandmestikel ning taotlustel. Mõlemad meetodid on sisseehitatud Texta Toolkitti¹⁶.

Tabel 10. BERT-mudelite treenimisel kasutatud parameetrid.

Eps	1.00E-08	Balance to max limit	FALSE
Learning Rate	0.00002	Number of Epochs	2
Minimum Sample Size	1	Negative Multiplier	1
Maximum Sample Size	10000	Split Ratio	0.8
Balance	FALSE	Max Length	64
Use sentence shuffle	FALSE	Batch size	32

¹⁶ <https://texta.ee/products#toolkit> - Texta Toolkit, vabavaraline tööriist tekstianalüütika jaoks

BERTi puhul on sisendiks tekstid märgenditega ning suur keelespetsiifiline mudel, näiteks eesti keele puhul EstBERT¹⁷ või inglise keele puhul *bert-base-uncased*¹⁸. BERT mudeleid treeniti ühe valdkonnaga (põhivaldkondade) treenimisandmete põhjal, kuna tavaliselt tuleb BERT sellega paremini toime. Tagger Groupi puhul oli sisendiks tekstid märgenditega, kuid üks tekst võib seostuda mitme valdkonnaga.

Tabel 11. Tagger Group mudelite treenimisel kasutatud parameetrid.

	Eesti keeles treenitud parim mudel	inglise keeles treenitud parim mudel
Balance	FALSE	FALSE
Balance to max limit	FALSE	FALSE
Analyzer	word	word
Vectorizer	Count Vectorizer	Hashing Vectorizer
Classifier	LinearSVC	Logistic Regression
Ignore numbers	TRUE	TRUE
Minimum sample size	25	25

Mõlema meetodi puhul kahaneb näidete arvu vähenemisega paratamatult ka tulemuste kvaliteet. Katsetati eri piirmääradega, kui mitu dokumenti peab ühe valdkonnaga seotud olema, et selle põhjal saaks midagi õppida. Texta Toolkit võimaldab mõlema meetodi puhul klasse tasakaalustada, kuid teeb seda tehnikult vähemesindatud klasside näiteid dubleerides ning tulemusi see ei parandanud. Tagger Groupi puhul võib samade parameetritega uut mudelit treenides saada teisgusused tulemuses, kuna enne treenimist sooritatakse dokumentide automaatne jaotus treening- ning testgruppi mõjutab edasist treenimist.

Generatiivsed suured keelemudelid (LLM-mudelid) on suutlikud ilma lisatreenimiseta inglise keeles häid tulemusi saavutama, samas on eesti keeles nende mudelite generatiivsed oskused kindlasti halvemad kui inglise keeles. LLM mudelid on lihtne lahendus, kuid nendega kaasnevad riskid nagu hallutsineerimine (ehk valeinfo loomine) ning majanduslike kulude kasv. Valdkondade määramise analüüsis kasutati OpenAI ChatGPT/GPT-4o mudelit.

¹⁷ <https://huggingface.co/tartuNLP/EstBERT> - EstBERT, eestikeelne BERT keelemudel

¹⁸ <https://huggingface.co/google-bert/bert-base-uncased> - bert-base-uncased, ingliskeelne BERT keelemudel

5.4 Meetodite analüüs, eksperimendid ja tulemused

Tulemused on esitatud testandmestikul ning taotlustel, mida mudeli treenimiseks ei kasutatud.

Testandmestikud on esitatud **tabelis 9**. Taotluste puhul võrreldi mudeli tulemusi, kui sisendiks oli eestikeelne annotatsioon, inglise keele puhul oli ennustusi rohkem, olenevalt sisendtekstist, kas taotluse annotatsioon, teaduslik põhjendus või projekti põhieesmärgid. Taotlusi oli 9.

Kvaliteedimõõdikud on esitatud kõigi valdkondade (klasside) peale kokku, mis tingib vajaduse arvutada klasside keskmise tulemuse kõigi mõõdikute kohta. Selles töös esitatakse kõik tulemused kasutades algoritmi *weighted*, mis võtab arvesse seda, et osa valdkondi esineb rohkem kui teisi - see annab tegelikkusele rohkem vastava tulemuse.

5.4.1 BERT-klassifitseerija abil valdkondade määramine

Tabel 12. BERT mudelite kvaliteedimõõdikud keele alusel, *weighted*-keskmine.

Mudeli ID	Keel	Testandmestiku nimetus	F1	Täpsus	Saagis
604	eesti	<i>P_secondary</i>	0.01	0.01	0.08
603	inglise	<i>OA_secondary</i>	0.06	0.05	0.15

Väga paljude klasside korruga määramisel tekib probleem, et mudel õpib ainult kõige sagedamini esinevaid klasse määrama ning seega määratakse kõik sissetulevad dokumendid sagedamini esinevatesse klassidesse (**tabel 12**). Näiteks mudeli 603 puhul on saagis suur kõigil klassidel, mida esineb treeningandmestikus rohkem kui 1500 korda, samas on sealne täpsus hästi madal (**tabel 13**).

Tabel 13. Sagedamini esinenud klasside kvaliteedimõõdikud BERT mudelil 603.

Valdkond	Täpsus	Saagis	F1
<i>Electrical and Electronic Engineering</i>	0.231	0.749	0.353
<i>Molecular Biology</i>	0.155	0.816	0.26
<i>Sociology and Political Science</i>	0.101	0.766	0.178

Kokkuvõtvalt leiti, et BERT klassifitseerija tulemused on halvad ning läksid veelgi halvemaks, kui dokumendiklasse jurde panna, mistõttu ei ole lähenemine sobilik sellise ülesande jaoks.

5.4.2 Tagger Group mudelite abil valdkondade määramine

Parimate Tagger Group mudelite sisendiks olid valdkonnad, millele vastas vähemalt 25 dokumenti. See tähendab, et inglise keeles sai 241 valdkonnast mudeli treenida 206 põhjal ning eesti keeles 319 valdkonnast 167 põhjal. Väiksema klassikaetuse tõttu on ka eesti keele mudeli tulemused testandmestikul kehvemad. Mudel ennustas vaikumisi kõiki valdkondi, mille seotus sisendiga oli vähemalt 50% tõenäoline, mudeli määratud valdkondade arv piirati viiele.

Tabel 14. Tagger Group mudelite tulemused testandmestikul, *weighted*-keskmine.

Mudeli ID	Keel	Testandmestiku nimetus	F1	Täpsus	Saagis
352	inglise	<i>OA_secondary</i>	0.44	0.36	0.62
351	eesti	<i>P_secondary</i>	0.23	0.16	0.46

Tulemused testandmestikel on üsna halvad, eriti eestikeelsed tulemused (**tabel 14**). Samas on tulemused märksa paremad kui BERT mudelitel. Ebavõrdne andmete jaotus on ilmselt suurim halbade tulemuste põhjustaja. Lisaks, lähedaste valdkondade sõnavara kattub suurel määral ning võimalikke klasse (valdkondi) on hästi palju.

Tabel 15. Tagger Group tulemused taotlustel, *weighted*-keskmine.

Mudeli ID	Keel	Testandmestiku nimetus	Väli	F1	Täpsus	Saagis
352	inglise	<i>Application</i>	Teaduslik taust	0.36	0.32	0.46
352	inglise	<i>Application</i>	Projekti eesmärgid	0.36	0.32	0.46
352	inglise	<i>Application</i>	Annotatsioon	0.42	0.41	0.46
351	eesti	<i>Application</i>	Annotatsioon	0.56	0.54	0.58

Mudelite keskmised kvaliteedimõõdikud pole kõrged (**tabel 15**), kuid arvestades suurt klasside arvu on tulemus ikkagi üsna täpne. Kindlasti mängib saadud tulemuste puhul rolli ka see, et taotluste arv on hästi väike, vaid üheksa. Valede valdkondade määramise põhjuseks on valdkondade omavaheline kattumine. On märgata, et eri väljade puhul on ingliskeelsel mudelil eri kvaliteet, see tuleneb ilmselt väljas oleva teksti sisulisest sarnasusest treenimisväljaga, parima tulemuse sai annotatsioon.

Lõplikeks mudeliteks treeniti Tagger Group mudelid kõigi olemasolevate ingliskeelsete ja eestikeelsete andmete põhjal, et tulemusi veidi parandada ning rohkem valdkondi katta. Eesti keeles treeniti mudel 188 klassiga ning inglise keeles 212 klassiga. Tulemused valideeriti taotluste peal (**tabelis 16**).

Tabel 16. Valitud Tagger Group mudelite kvaliteedimõõdikud taotlustel, *weighted*-keskmine.

Mudeli ID	Keel	Testandmestiku nimetus	Väli	F1	Täpsus	Saagis
360	inglise	<i>Application</i>	Teaduslik taust	0.38	0.36	0.46
360	inglise	<i>Application</i>	Projekti eesmärgid	0.36	0.32	0.46
360	inglise	<i>Application</i>	Annotatsioon	0.40	0.37	0.46
356	eesti	<i>Application</i>	Annotatsioon	0.81	0.79	0.83

Parim eestikeelne mudel (mudel 356 **tabelis 16**) on CERCS koodide ja eestikeelsete CERCS nimetustega. Selle F1 tõusis 25 protsendi võrra ning saagis ja täpsus paranesid. Üheksast taotlusest seitse said õige valdkonna. Kuna kvaliteedi tõus on järsk, siis ei saa välistada, et statistiliselt õppis mudel määrama nimelt neid valdkondi, mis olid taotlustes esindatud.

Ingliskeelsete andmete puhul pole testandmestikust andmete lisamisel paranemist näha, seega tuleks uurida veel sisendandmete kvaliteedi parandamise võimalusi ning samuti saaks veel

korrigeerida klasside tunnuseid. Samuti võib uurida, kui suur peab olema valdkonna määramise tõenäosus, et seda määratakse.

5.4.2.1 OpenAlex põhivaldkonna pealt treenitud Tagger Group

Treeniti uus mudel, mis võttis sisendiks lühikokkuvõtted, kuid sisendvaldkondadeks olid OpenAlexi määratud põhivaldkonnad (*primary subfield*). Tulemused on allpool tabelis esitatud taotluste testandmestiku pealt arvutatud keskmistena.

Tabel 17. OpenAlex põhivaldkonna pealt treenitud Tagger Group mudelid tulemused taotlustel, *weighted*-keskmisega.

Mudeli ID	Keel	Testandmestiku nimetus	Väli	F1	Täpsus	Saagis
361	inglise	<i>Application</i>	Teaduslik taust	0.50	0.49	0.54
361	inglise	<i>Application</i>	Projekti eesmärgid	0.38	0.35	0.46
361	inglise	<i>Application</i>	Annotatsioon	0.38	0.38	0.38
361	inglise	<i>Application</i>	Teaduslik taust Projekti eesmärgid Annotatsioon	0.38	0.35	0.46

Tulemused olid veidi paremad kui eelnevalt treenitud mudelitel, kõrgeim F1 on 0.5 teadusliku tausta juures (**tabel 17**). Annotatsioonide puhul siiski olid tulemused pisut halvemad. Kasutati ka kõiki ingliskeelseid välju, et sisendteksti oleks rohkem, aga ka sellest ei tulnud paremaid tulemusi, tulemused jäid projekti eesmärgi tasemele.

5.4.2.2 Ühe valdkonna määramine Tagger Group abil suurema andmestiku põhjal

Klassifitseerimise katseks võeti üks valdkond – *Parasitology* (parasitoloogia). OpenAlexi API kaudu hõivati 10 000 antud valdkonna dokumenti. Lisandunud andmemahust olid sisukad 9354 dokumenti, millest parasitoloogia põhivaldkonnaga oli 4752. Andmestikule lisati negatiivsete näidetena ka 10 000 mitte-parasitoloogiaga seotud dokumenti varasemalt kogutud OpenAlex andmete seast.

Testandmestikuks olid OpenAlexist juba varem kraabitud andmed, mille põhivaldkonnaks oli määratud parasitoloogia (71 tükki). Dokumentidele määratud valdkonna ja sisu vastavus kontrolliti käsitsi üle, et kontrollandmete kvaliteeti parandada. Parasitoloogia valdkonna valik tulenes sellest, et valdkond on üsna selgepiiriline ning ka valdkonnakaue inimene saab selle valdkonna tekstidest otsimootori abiga aru ning suudab määrata, kas tekst seostub parasiitide ja parasiitide tekitatud/levitatud haigustega. Käsitsi määrares oli 71st dokumendist 65 parasitoloogiaga seotud.

Nagu näha, siis mudelite tulemused paranesid võrreldes varasemaga (**tabel 18**). Sooritati ka võrdlus OpenAlexi korrigeerimata andmetel ehk OpenAlexi automaatselt määratud põhivaldkondadel. Mudelid peaksid sel juhul kõik 71 testdokumenti määrama parasitoloogia valdkonda. Eksperiment oli oluline, sest automaatselt määratud valdkondadega testandmestik annaks paratamatult teisi tulemusi kui käsitsi määratud valdkondadega.

Tabel 18. Parasitoloogia Tagger-mudelite kvaliteedimõõdikute tulemused, arvutusalgoritm *weighted*.

Võrreldavate valdkondade kirjeldus	Testandmestiku nimetus	F1	Täpsus	Saagis
Käsitsi korrigeeritud valdkonnad	<i>Parasite_primary</i>	0.88	0.90	0.87
Käsitsi korrigeeritud valdkonnad	<i>Parasite_secondary</i>	0.88	0.87	0.89
OpenAlexi korrigeerimata valdkonnad	<i>Parasite_primary</i>	0.93	1.00	0.87
OpenAlexi korrigeerimata valdkonnad	<i>Parasite_secondary</i>	0.97	1.00	0.94

Käsitsi määratud andmete täpsus on *weighted*-algoritmi järgi sama või veidi madalam kui OpenAlexi automaatsete valdkondade puhul. See on seetõttu, et OpenAlexi andmestik on üleüldiselt veidi vigane ning nende andmete põhjal treenitud mudelitesse jääb see ebatäpsus sisse.

Kui võrrelda *weighted*-algoritmiga saadud uute mudelite tulemusi eelnevatega, siis on tulemused head. Suurema andmemahuga on kvaliteedimõõdikud paremad kui enne, kuid võib oletada, et tulemused lähevad veidi halvemaks, kui rohkem valdkondi sisse tuua ja keskmise arvutamine muutub keerulisemaks.

5.4.2.3 Kümne valdkonna määramine Tagger Group abil suurema andmestiku põhjal

Kuna ühe valdkonna puhul kaasnes andmete lisamisega parem kvaliteet, siis laiendati valdkondade arvu, et saaks parema ülevaate andmekogu suurendamise mõjust. Seekord kaasati OpenAlexist rohkem annotatsioone kümne vaimse tervisega seonduvast valdkonnast: *Psychiatry and Mental health, Social Psychology, Clinical Psychology, Experimental and Cognitive Psychology, General Health Professions, Developmental and Educational Psychology, Behavioral Neuroscience, Applied Psychology, Health, Biological Psychiatry*. Valdkondade valik põhineb eeldusel, et neid valdkondi on keerulisem selgepiirilisel määra. Samuti võeti arvesse valdkondade esinemissagedus algandmete seas.

Igast valdkonnast koguti 10 000 dokumenti (kokku 100 000), dokumendid puhastati sarnaselt varasematele andmetele. 73 696 dokumenti jäi puhastuse järel alles, neist 41 757 dokumendis oli üks eelpool loetletud klassidest põhivaldkonnaks.

35 982 annotatsiooni ning põhivaldkonna põhjal treeniti Tagger Group mudel. Testandmestik moodustati uutest andmetest, igast klassist kaasati vähemalt 700 dokumenti, mida ei esinenud algselt kogutud andmetes. Vaimse tervise põhivaldkondadega testandmestik oli 5697 dokumendi suurune. 32 017 dokumenti (43,5% kogutud andmetest) jäi analüüsis kasutamata vale põhivaldkonna tõttu.

Võrdluseks treeniti algselt kogutud andmetel samuti uus Tagger Group mudel, mis hõlmas vaid eespool loetletud valdkondi. Mudel treeniti 2235 dokumendi põhjal ning andmetes esines tõsine tasakaalu puudus – mõnda valdkonda kuulus üle 500 dokumendi, teist esindas aga vaid 17 dokumenti.

Tabel 19. Vaimse tervisega seonduvate klasside tulemused nii uue mudel (lisatud andmetega) kui vana mudeli põhjal. Võrdluseks on põhivaldkond, *weighted*-keskmine.

Valdkonna nimetus	Lisatud andmetega Tagger Group mudeli tulemused (<i>Psych_new</i>)			Lisandmeteta Tagger Group mudeli tulemused (<i>Psych_old</i>)		
	Täpsus	Saagis	F1	Täpsus	Saagis	F1
Applied Psychology	0.13	0.92	0.23	0.13	0.55	0.21
Behavioral Neuroscience	0.44	0.93	0.60	0.42	0.69	0.52
Biological Psychiatry	0.43	0.95	0.59	0.27	0.82	0.40
Clinical Psychology	0.31	0.87	0.45	0.21	0.72	0.32
Developmental and Educational Psychology	0.28	0.86	0.42	0.16	0.52	0.24
Experimental and Cognitive Psychology	0.14	0.83	0.23	0.11	0.60	0.19
General Health Professions	0.20	0.82	0.33	0.12	0.64	0.20
Health	0.31	0.91	0.47	0.17	0.59	0.26
Psychiatry and Mental health	0.67	0.81	0.74	0.38	0.78	0.51
Social Psychology	0.42	0.81	0.55	0.37	0.61	0.46
Kõik valdkonnad kokku	0.42	0.85	0.54	0.29	0.68	0.39

Kuigi lisatud on rohkem kui kümme korda enam dokumente, siis tulemused põhivaldkondade määramisel oluliselt ei parane (**tabel 19**). Kõigi kümne valdkonna peale on küll nii täpsus kui saagis tõusnud rohkem kui 10 protsenti, kuid F1 jääb siiski 0.54 ligi, mis pole piisavalt hea, et mudeleid

kasutusele võtta. Ainus valdkond, mida saaks antud mudeliga mingil määral määrata oleks *Psychiatry and Mental health*.

Kui otsida valdkondi kõigi dokumendile määratud valdkondade seast, siis on situatsioon sarnane (**tabel 20**). Kuigi võiks eeldada, et kõigi valdkondade hulgas oleks mudelil rohkem võimalusi korrektselt valdkonda määrata, siis jääb F1 0.51 tasemele, ehk siis saaks selle mudeli abil määrata õigesti ainult 51% kõigist valdkondadest. Nagu varasemas **tabelis 19**, siis ka sellel puhul suudaks mudel võib-olla määrata valdkonda *Psychiatry and Mental health* ja võib-olla valdkonda *Social Psychology*.

Tabel 20. Vaimse tervisega seonduvate klasside tulemused nii uue mudel (lisatud andmetega) kui vana mudeli põhjal. Võrdluseks on kõik valdkonnad, *weighted*-keskmine.

Valdkonna nimetus	Lisatud andmetega Tagger Group mudeli tulemused (<i>Psych_new</i>)			Lisaandmeteta Tagger Group mudeli tulemused (<i>Psych_old</i>)		
	Täpsus	Saagis	F1	Täpsus	Saagis	F1
Applied Psychology	0.35	0.82	0.49	0.28	0.39	0.33
Behavioral Neuroscience	0.64	0.68	0.66	0.68	0.57	0.62
Biological Psychiatry	0.69	0.68	0.69	0.45	0.61	0.52
Clinical Psychology	0.54	0.61	0.58	0.42	0.58	0.49
Developmental and Educational Psychology	0.49	0.64	0.56	0.32	0.44	0.37
Experimental and Cognitive Psychology	0.35	0.64	0.45	0.31	0.49	0.38
General Health Professions	0.45	0.59	0.51	0.29	0.55	0.38
Health	0.56	0.71	0.63	0.32	0.49	0.39
Psychiatry and Mental health	0.88	0.62	0.72	0.59	0.71	0.64
Social Psychology	0.78	0.64	0.70	0.70	0.49	0.57
Kõik valdkonnad kokku	0.52	0.52	0.51	0.4	0.45	0.41

Antud katse tõestab, et isegi kui andmeid lisada, siis on valdkondade määramise probleem keerulisem kui alguses tundub. Samu probleeme uurivad valdkonnad kasutavad sarnast sõnavara ning sarnaseid meetodeid. Seetõttu on valdkondi keeruline ainult annotatsiooni alusel eristada. Samamoodi mõjutavad saadud tulemusi tõenäoliselt OpenAlexi ebausaldusväärsed valdkonnamääratlused, mida pole küll palju, kuid mis takistavad mudeleid valdkondade eristamisel.

5.4.3 LLM-i abil valdkondade määramine

Lisaks klassikalisele masinõppele katsetati ka seda, kas suured keelemudelid on võimelised valdkondi automaatselt ennustama. Selleks valiti kogutud andmetest juhuslikkuse alusel kakskümmend publikatsiooni lühikokkuvõtet koos OpenAlexi valdkondadega ja kõigi andmestikus esinevate valdkondade nimekirjaga. Valdkondade nimekiri ja lühikokkuvõtted edastati kolm korda ChatGPT-le¹⁹ valdkonna määramiseks, et võrrelda, kui palju pakutud valdkonnad varieeruvad. Vaikimisi seatud parameetreid ei muudetud. Määramisjuhend oli järgnev:

"You are a specialist who assigns fields to scientific publications. The list of all possible fields is given here: [...].

Please carefully read the content of the document and classify it into the most appropriate field from the list above. Provide ONLY one field, do NOT provide any explanation nor additional text.

Document: [...]

25% (kahekümnest etteantud dokumendist viies dokumendis) on kolme ennustuse hulgas olemas õige põhivaldkond (*primary subfield*). 55% juhtudest (üheteistkümmel dokumendil) on ennustuste hulgas kõikide valdkondade (*secondary subfield*) alla kuuluv valdkond, sealhulgas need viis dokumenti, kus oli ka õige põhivaldkond.

Üheksa dokumendi puhul ehk alla poolte kordade olid kõik kolm ennustust täpselt samad. Vaid ühel korral ennustas ChatGPT ühele dokumendile iga kord erineva valdkonna. Ülejäänud dokumentidele ennustas ChatGPT vähemalt kahte erinevat valdkonda.

Mudel kipub vahel pakkuma ka etteantud valdkondade hulka mitte kuuluvaid klasse – seda üheksa ennustuse puhul (viie erineva dokumendi raames). Nendest enim kaldus mudel ennustama valdkonda *Management*, mida nimekirjas polnud. Seega eiras mudel juhendit, et oli vaja valida sobivaim valdkond saadetud nimekirjast.

Tulemused täispikkuses on **lisas 1**, kus on eraldi märgitud väljamõeldud valdkonnad helepunase värviga, vähemalt korra õigesti määratud valdkonnad helerohelise värviga ning ChatGPT õigesti määratud, kuid OpenAlexis teise valdkonnaga kollase värviga. ChatGPT-d kasutades ei ole suuremat kattuvust OpenAlexi määratud valdkondadega, kahel juhul OpenAlexi valesti määratud valdkonna tõttu, näiteks https://doi.org/10.1163/9789004545687_004 puhul tundub kokkuvõtte religiooniajaloo valdkonda kuuluvat nagu ChatGPT on määranud. Samamoodi tundub [dx.doi.org/10.1515/fca-2017-0058](https://doi.org/10.1515/fca-2017-0058) kuuluvat valdkonda *Numerical Analysis*. Mõnel juhul on valdkonnad väga sarnased, mis põhjustab valesti määramist (näiteks *Signal Processing* ja *Computer Vision and Pattern Recognition*).

Kui lisada kaks ChatGPT inimhinnanguga ühtivat valdkonda õigete ennustuste hulka, siis jääb ikkagi seitsmel dokumendil õige valdkond määramata (ehk siis vähemalt ühe õige valdkonna määramise

¹⁹ <https://openai.com/chatgpt/> - ChatGPT, suur keelemudel.

osakaal jääks umbes 65% juurde). LLM-i kaudu valdkondade määramise tulemused on sarnased Tagger Groupiga saadud tulemusega, kuid pole usaldusväärsed, kui arvestada väljamõeldud valdkondi ning ka üsna suurt varieeruvust valdkondade määramisel. Lisaks tuleb arvestada GPT kasutamisega kaasnevat päringute kulu ning autoriõigustega seotud probleemide vältimiseks uurida õiguslikku tausta (näiteks OpenAlexis pole tegelikult kokkuvõtted automaatselt teksti kujul kättesaadavad²⁰).

5.4.3.1 Teiste päringute mõju LLM-i määratud valdkondadele

Samade annotatsioonide põhjal katsetati uuesti LLM abil valdkonna määramist, seekord kasutati kahte uut päringut. Ühes lisandunud päringus olid ühendatud erinevad lähenemised tulemuste üldiseks parandamiseks²¹ (näiteks *chain-of-thoughts* meetod) ja teises oli lisatud näide parasitoloogia valdkonna annotatsioonist ning sellele määratud valdkonnast. Kui eelmises katses kasutati keelemudelina ChatGPT-d, siis seekord kasutati OpenAI GPT-4o²² mudelit.

Teine päring:

"You are Frederick, an AI expert in assigning fields to scientific publications. You are tasked with sorting through scientific publications by analysing their content and deciding which field they belong to."

'Yes, I understand. I am Frederick, and I will analyse the scientific publication.'

'Great! Let's begin then :)

For the given abstract: [...]

What is the field that this abstract belongs to? Respond with the field only please.

Answer: Let's work this out in a step by step way to be sure we have the right answer."

Kolmas päring:

"You are an AI expert in assigning fields to scientific publications. You are tasked with sorting through scientific publications by analysing their content and deciding which field they belong to.

The list of all possible fields is given here: [...].

Example:

Q: The molecular characterisation of Giardia and Cryptosporidium has given rise to a more epidemiological meaningful and robust taxonomy. Importantly, molecular tools are now available for 'typing' isolates of the parasites directly from clinical and environmental samples. As a consequence, information on zoonotic potential has been obtained although the frequency of zoonotic transmission is still poorly understood. Analysis of outbreaks and case-control studies, especially when coupled with genotyping data, is slowly providing information on the public health significance of zoonotic transmission. Such studies support the hypothesis that Cryptosporidium hominis is spread only between humans but that the major reservoir for Cryptosporidium parvum is domestic livestock, predominantly cattle, and

²⁰ https://docs.openalex.org/api-entities/works/work-object#abstract_inverted_index - OpenAlex, info sisukokkuvõtete kohta

²¹ Clavié; Benjamin jt 2023. Large Language Models in the Workplace: A Case Study on Prompt Engineering for Job Type Classification. NLDB 2023

²² <https://openai.com/index/hello-gpt-4o/> - GPT-4o, suur keelemudel

that direct contact with infected cattle is a major transmission pathway along with indirect transmission through drinking water. The situation is less clearcut for Giardia duodenalis but the evidence does not, in general, support zoonotic transmission as a major risk for human infections. However, for both parasites there is a need for molecular epidemiological studies to be undertaken in well-defined foci of transmission in order to fully determine the frequency and importance of zoonotic transmission.

A: Parasitology

Please carefully read the content of the document. Classify it into the most appropriate field from the list above and return only ONE field with no explanation. Document: [...]

Õige valdkonna ennustamine oli esimese päringuga võrreldes vähem edukas. Teise päringuga leiti kolme ennustuse hulgast õige põhivaldkond kahe dokumendi puhul (10%), kolmanda päringuga nelja dokumendi puhul (20%). Teise päringuga leiti kokku kõikide valdkondade (*secondary subfield*) alla kuuluv valdkond viie dokumendi puhul (25%), kolmanda päringuga kümne puhul (50%). Esimesel päringul olid need näitajad 25% ja 55%.

Teine päring soodustas valdkondade väljamõtlemist, valdav enamik ennustatud valdkondi puudusid etteantud nimekirjast (47 ennustatud valdkonda 60st). See võis tuleneda sellest, et päringus polnud LLM-ile otseselt selgitatud, et ülesandeks on valdkond määrata etteantud nimekirja hulgast. Kolmas päring parandas tulemusi vähemalt sedavõrd, et LLMi leiutatud valdkondi oli vähem, vaid kaks tükki.

Teise päringu kasutamisel kattusid kõik kolm ennustust 35% dokumentide puhul, 25% dokumentidest olid aga saanud endale kolm erinevat valdkonda. Kolmanda päringu puhul olid tulemused veidi paremad, 60% dokumentidest said mudeli järgi sama valdkonna, kolm erinevat valdkonda said 15% dokumentidest.

Kui päringusse lisada näide valdkonna määramise kohta, siis vähenesid väljamõeldud valdkonnad ning hästi varieeruvad ennustused, aga vaid pooltele dokumentidele määrati õige valdkond.

Tulemused täies mahus on [lisas 2](#) ja [lisas 3](#). Tabelites on eraldi märgitud väljamõeldud valdkonnad helepunase värviga, vähemalt korra õigesti määratud põhivaldkonnad helerohelise värviga ning kõigi valdkondade seast määratud õiged valdkonnad kollase värviga.

5.5 Valdkonna määramise tulemuste kokkuvõte ja kontseptuaalne prototüüp

Projektide annotatsioonidel treenitud Tagger Group mudel, mis määras 81% taotlused õigestesse valdkondadesse, võiks töötada tekstidele valdkondade määramiseks. OpenAlexi andmetel treenitud mudelite sooritus jäi liiga kehvaks, et neid kasutada valdkondade automaatseks määramiseks.

BERT klassifitseerimismudel ei sobi antud ülesande lahendamiseks, kuna klasse on liiga palju. Ainus variant oleks treenida igale valdkonnale erinev binaarne mudel, mis määrab, kas sisendtekst kuulub antud valdkonda või mitte.

LLM-mudelite tase jäi samaväärseks OpenAlex andmetel põhinevate Tagger Group mudelite tulemustega. Tagger Groupile aga jäi antud variant alla, sest LLM-i määratud valdkonnad ei olnud usaldusväärsed, kuna korduval küsimisel määrati samadele annotatsioonidele erinevaid valdkondi. Samuti pakkus LLM välja uusi valdkondi, mida süsteemile sisendiks ei antud ning mis valdkondade hierarhiast puudusid.

5.6 Lisategevused tulemuste parandamiseks

Kui oleks võimalik OpenAlexi vähema andmemahuga valdkondadesse lisada andmeid juurde, siis see annaks ilmselt kõige parema tulemuse. Samas on risk, et andmete hulk läheb liiga suureks ning andmetes on jätkuvalt ebatäpsused, mis mõjutavad ka määratavat valdkonda. OpenAlexi andmete puhul tuleks uurida veel sisendandmete kvaliteedi parandamise võimalusi, näiteks lühemate tekstide eemaldamist, samuti saaks veel korrigeerida klasside tunnuseid. OpenAlexi ebatäpsuste vähendamiseks tasuks andmeid märgendada, see tähendab, et käsitsi määrata suuremale hulgale dokumentidele õige valdkond. Võimalus oleks veel kasutada üldisemaid valdkondi, kuid see oleneb sellest, kas üldisem valdkond sobiks taotluste valdkonna määramiseks. Iseenesest aitaks see selle vastu, et klasse on hästi palju ning eemaldaks võimaluse, et mudel määrab tekstile sarnase, kuid mitte täpse valdkonna.

ETISe andmeid aluseks võttes oleks mõistlik testida, milline oleks Tagger Groupi mudelite tulemus, kui treenimiseks kasutada projektide ingliskeelseid annotatsioone ning CERCS valdkondi.

Tabel 21. Ülesanded, mida võiks tulevikus teostada, koos ülesande mahuhinnanguga.

Andmestik, millega jätkata (OpenAlex lisaandmed või ETIS projektide ingliskeelsed annotatsioonid)	Ülesande kirjeldus	Eeldatav maht
Kumbki andmestik	Filtreerida välja ainult metaandmeid sisaldavad sisukokkuvõtted regulaaravaldistega	30 tundi
Kumbki andmestik	Koodi või sümboleid sisaldavate kirjete puhastamine regulaaravaldistega	40 tundi
OpenAlex	Lisaandmete hõivamine	40–60 tundi (olenevalt andmemahust)
OpenAlex	Valdkondade käsitsi märgendamine	150–200 tundi (olenevalt andmemahust)
ETIS projektide ingliskeelsed annotatsioonid	Andmete töötlemine	15 tundi
ETIS projektide ingliskeelsed annotatsioonid	Tagger Mudelite treenimine ja valideerimine	30 tundi
Kumbki andmestik	Valdkondade tunnuste korrigeerimine (stoppsõnade lisamine Tagger Group mudelitele)	30–45 tundi (olenevalt andmemahust)

6. Arengukavade vastavusanalüüsi abilise analüüs

Arengukavadest vastete leidmine abistab taotlustest automaatselt leida vastavusi arengukavades kirjeldatud probleemidega või seatud eesmärkidega, samuti üldisemalt, millise arengukava teemadega taotluse tekst haakub.

6.1 Alusandmed ning nende töötlemine

Alusandmeteks oli Teadusagentuurilt saadud 34 arengukava sisu, mis parsiti teksti kujule ning mis rikastati algdokumentidest eraldatud metaandmetega – andmeallika ning dokumendi nimetusega. Arengukavad parsiti teksti kujule *pypdf*²³ teegiga, mis on Pythoni teek PDF-dokumentide töötlemiseks. Üks *docx*-kujul fail muudeti parsimiseks PDF kujule, sest arengukavad segmenditi lehekülje kaupa.

Segmenteerimine on vajalik otsingus kasutatavate vektormudelite sisendi pikkuse piirangute tõttu ning ka täpsema vaste leidmiseks, et kasutaja ei peaks vasteid pikemast tekstihulgast otsima. Lehekülje kaupa segmenteerimisel jääb alles palju konteksti võrreldes lõigu või lausega ning lehekülge on lihtne määratleda. Leheküljed eraldati PDF dokumendis lehekülje-piiri kasutades. Kokku oli lehekülgi ehk segmente 1502, millest peale puhastamist jäi alles 1442. Andmestikus segmentides esinenud sõnade keskmine arv oli 315. Iga segmendi sisu algusesse lisati metaandmed, et allikas ning dokumendi nimi annaks veidi rohkem kontekst. All on toodud näide ühe segmendi algusest:

*Kultuuriministeerium
Kultuuri arengukava
10*

*2. Eesti kultuurimälu on hoitud ja hinnatud
Eesti rikkalik vaimne ja aineiline kultuuripärand on hästi hoitud, uuritud ja teadvustatud,
kultuuripärand on elavas kasutuses ja inimestele oluline, tagatud on kultuuripärandi
pikaajaline
füüsiline ja digitaalne säilimine ning kättesaadavus.*

²³ <https://pypdf.readthedocs.io/en/stable/> - PyPDF, Pythoni teek PDF-dokumentide töötlemiseks

6.2 Arengukavadele vastavuse leidmise meetodid

Vastavuse leidmiseks andmed vektoriseeriti ning vektoriseeritud andmetele sooritati päringuid kahe sarnasusotsingute (ehk vektorotsingute) sooritamise meetodi abil. Järgnevalt pikemalt vektoriseerimisprotsessist ning vektorotsingute liikidest.

6.2.1 Andmete viimine vektorkujule

Vektorotsingud võimaldavad hinnata andmebaasis olevate dokumentide sarnasust sisendile, mitte ainult sõnade ja fraaside tasandil, vaid ka sarnasusi sisus. Antud projektis vektoriseeriti dokumentide segmendid kasutades Pekingi Tehisintellektiakadeemia vabavaralist vektormudelit BGE-M3²⁴, mis varasemate eksperimentide baasil on andnud eesti keelel häid tulemusi RAG (*retrieval augmented generation*) eksperimendis võrreldes Azure AI ja Ada vektoritega seadustest vastete leidmisel. Vektormudel genereerib igale segmendile sellele vastava tekstivektori. Vektoriseerimiseks kasutati ühte GPU-d, sarnase riistvarata on protsess aeglasem või mahukama andmestiku puhul sisuliselt võimatu. Näide vektoriseerimisest:

```
Sisend tekstina: metsanduse arengukava
Tekstivektor: [0.36688125133514404, 0.5853566527366638, 0.519641637802124,
-0.05295291915535927,...]
```

Lisaks arengukavade vektoriseeritud tekstile on iga segmendi juures vastava asutuse või allika nimi, algse dokumendi nimi (enamikel juhtudel arengukava pealkiri), lehekülje number ning lehekülje sisu. Andmed on otsingute sooritamiseks indekseeritud Elasticsearch otsingumootoris.

6.2.2 Sarnasusotsingute liigid

Sarnasusotsingud viitavad otsingutele, kus võrreldakse tekste ning leitakse sarnaseim tekst. Vektorotsingutes võrreldakse tekstivektoreid, arvutades kahe vektori vahelise kauguse, kuid erinevaid võimalikke arvutusalgoritme on palju.

Analüüsi koostamiseks kasutati kahte liiki sarnasusotsinguid, *k-nearest neighbour*²⁵ (edaspidi **KNN**) ja Elasticsearchi More Like This²⁶ (eesti keeles "rohkem sarnaseid dokumente"; edaspidi **MLT**).

MLT võimaldab otsimootorilt pärida sarnaseid dokumente kõige olulisemate võtmesõnade alusel. KNN on vektorotsing ning leiab sisendile sarnasemate dokumentide grupi. Kui MLT päringu sisendiks on tekst, siis KNNi päringu sisend tuleb viia vektorkujule. Mõlemat sarnasusotsingu liiki kasutati Elasticsearchi päringu kaudu.

²⁴ <https://huggingface.co/BAAI/bge-m3> - BGE-M3, vektoriseerimismudel HuggingFace keskkonnas

²⁵ https://en.wikipedia.org/wiki/K-nearest_neighbors_algorithm - *k-nearest neighbour* sarnasusotsingu kirjeldus

²⁶ https://www.elastic.co/guide/en/elasticsearch/reference/7.17/query-dsl-mlt-query.html#_how_it_works - Elasticsearch More Like This päringu kirjeldus

Kummagi otsinguliigiga leiti iga sisendi vasteks kümme teksti arengukavadest. Väljundite arvu saab muuta, kuid otsing annab vaikumisi alati täpselt väljundite arvu, ehk siis otsing tagastab seadistatud arvu dokumente ka info puudumisel või veelgi rohkemate vastete olemasolul.

Kuna annotatsioonid olid lühikesed, siis sai mõlemal juhul panna sisendiks terve annotatsiooni, kuid tuleb arvestada, et kummagi meetodi puhul ei saa taotlust täispikkuses sisendina kasutada, vaid ka see tuleks osadeks segmenteerida. KNN puhul on sisendi suurus väiksem kui MLT puhul.

6.2.3 Sisendid otsingutele

MLT ja KNN otsingut testiti üheksa taotluse pealkirjade ja annotatsioonidega. Samuti võeti testimiseks kümne projekti annotatsiooni, kus oli mainitud arengukava. Kõik tekstid, millest vastavusi otsiti, olid eesti keeles.

Kuna taotlustes ega projektide annotatsioonides ei esinenud väga palju arengukavade sisule viitavat teksti, siis moodustati veel lisaks kuus sünteetilist näidet. Näited moodustati, et kontrollida vektorotsingut juhtudel, kui otsingutekst on otse allikast, sõnastust on veidi muudetud, algallika teemat on laiendatud, otsingutekst on üldine/teisest žanrist, või otsinguteksti on oluliselt ümbersõnastatud (**tabel 22**).

Tabel 22. Sünteetiliste otsingute testnäited koos kirjelduste ja allikaga:

Näite sisu	Näite kirjeldus ja allikas
Destination Earth – tuntud ka kui DestinE – on Euroopa Komisjoni algatus, mille eesmärk on luua Maa digitaalne simulatsioon digitaalse kaksikuga, mida kasutatakse kliimamuutuste ja keskkonnakatastroofide mõju paremaks mõistmiseks ning võimaldamaks poliitikakujundajatel nendele probleemidele tõhusamalt reageerida,	Sõnasõnaline kattuvus (Keskkonnavaldkonna arengukava mustand)
Leitakse vastus küsimusele, kas arengukava on vaja ja miks? Üks peamistest põhjustest strateegilise plaani väljatöötamiseks on luua ühtne arusaam ja eesmärk organisatsiooni tegevustele.[2] Tähtis osa on organisatsiooni missiooni sõnastamisel. Missioon kirjeldab organisatsiooni eksisteerimise aluseid, aitab määratleda selle identiteeti ja sihtrühma. Missiooni sõnastamisel tuleb vastata järgmistele küsimustele: "Kes me oleme? Mille nimel? Kelle jaoks? Kus? Mida me teeme?" Missiooni sõnastus peab olema lühike, kirjutatud selges keeles, positiivne ja ligitõmbav, oleviku vormis ja kindlas	<u>Vikipeedia artiklist arengukavade kohta, üldine/teine žanr.</u>

<p>kõneviisis.[1] Näiteks Tartu Ülikooli arengukavas aastateks 2009–2015 on seatud Tartu Ülikooli missiooniks olla teadmistepõhise ühiskonna arengut juhtiv jõud ja järjepidevuse tagaja Eestis.[3]</p>	
<p>Sporditaristu peaks olema kõigile kättesaadav. Analüüsitakse spordiparkide, jalgpalliplatside ning muude rajatiste ehituse ning rahastuse protsesse, analüüsitakse linnaplaneerimisvalikuid, mis on tehtud, ning suhteid keskkonnaga ja rahvastikutihedusega.</p>	<p><i>Veidi muudetud sõnastusega ning laiendatud teemaga näide:</i> 5.6. Planeerimisel ja projekteerimisel käsitletakse liikumispaiku ning -alasi avaliku ruumilahutamatu osana. Sporditaristu peab olema kõigile kättesaadav. (Eesti spordipoliitika põhialused)</p>
<p>Kohalikke loodusvarasid vääripäädavad teadus- ning arendusorganisatsioonid, start-upid ning ärid jätkusuutlikult, elurikkust arvestavalt ning ressurssi hoides. Keskendutakse nii toormaterjalile kui jäätmetele, luues taaskasutuseks biogaasi ning sööta.</p>	<p><i>Tugevalt ümbersõnastatud:</i> Teadus- ja arendustegevuse, innovatsiooni ning ettevõtluse toel vääripäädakse kohalikke ressursse kestlikult, elurikkusega arvestavalt ja kõrge ressursitootlikkusega, keskendudes nii esmasele kui ka sekundaarsele toormele ning võimendades bio- ja ringmajandust. (TAIE fookusvaldkondade teemalehed)</p>
<p>Toetatakse jätkuvalt eestikeelsete raamatute ilmumist (sealhulgas tõlkimist), teatreid, kino, laule ja tantse ning kultuurset ajakirjandust.</p>	<p><i>Tugevalt ümbersõnastatud:</i> Jätkatakse eestikeelse kultuuri, sh kirjanduse, teatri, filmi, laulu- ja pärimuskultuuri, kultuuriajakirjanduse ning eesti keelde ja keelest tehtavate tõlgete toetamist. (Eesti keele arengukava)</p>
<p>Nooremad inimesed ei õpi enam oma kodukoha keelekasutust ega dialekte.</p>	<p><i>Väga tugevalt ümbersõnastatud:</i> Eesti keele murrete ja piirkondlike erikujude põlvkondadevaheline keeleülekanne on vähenenud. (Eesti keele arengukava)</p>

6.3. Arengukavadele vastavuse leidmise tulemused

Mõlemad sarnasusotsingu liigid leidsid iga sisendi vasteks kümme tõenäolisemat segmenti arengukavadest koos sarnasusskooriga (*score*). Sarnasusskoor võimaldab sarnasusotsingul saadud vasteid järjestada, see tähendab, et kõige sarnasem või tõenäolisem vaste on esimene. Sarnasusskoor erineb erinevate sarnasusotsingu liikide vahel, näiteks projektides oli MLT sarnasusskooride vahemik 8–22, KNNil aga 1,55–1,75.

Tagastatud segmendid vaadati üle ning nende vastavust sisendile analüüsiti käsitsi. Segmentide vastavust algsisendile hinnati skaalal hea, rahuldav või halb. Skaala punktide kriteeriumid ja näited on allpool:

Hea – leitud segmendid on kõik otsinguks kasutatud teksti spetsiifilisemate teemadega seotud (all tabelites rohelisega märgitud).

Sisend: *sihtrühmade ootused metsanduse arengukavale*

Väljund: *metsanduse arengukava segmendid, eriti need, mis puudutavad erinevate sihtrühmade ootusi*

Rahuldav – leitud segmentide seas on osa segmente sisendi spetsiifiliste teemadega seotud, kuid osa mitte; segmentides oli vaid üldisem sisendi teema (all tabelites kollasega märgitud).

Sisend: *loodushariduse edendamine õppekava täiendavate looduslaste õppeprogrammide läbiviimise kaudu õpilastele ja lastele*

Väljund: *mõned loodusharidusega seotud segmendid, üldisemad keskkonna- ja haridusega seotud segmendid*

Halb – ei leitud segmente, millel oleks seos sisendi teemadega (all tabelites punasega märgitud).

Sisend: *eesti kultuuripärandi hoidmise ja väärtustamise arengukava koostamine*

Väljund: *segmendid on hoopis teise teemaga, näiteks ehitus, kliima, põlevkivi jne, kultuuriteemat ei esine*

6.3.1 Taotluste annotatsioonide sarnasusotsingute tulemused

Taotluste annotatsioonide puhul olid tagastatud vasted üldiselt rahuldava kvaliteediga, kuna enamikel taotlustel olid üpris spetsiifilised teemad, mida arengukavades ei käsitletud. Esines ainult paar taotlust, nt saagi jätkusuutlikkuse või ökosüsteemide taastuvuse-teemalised, mille puhul oli võimalik siduda konkreetne teema mõne arengukavas esitatud teemaga, kuid mitte konkreetse punktiga. Seetõttu hinnati juba üldise teema leidmist heaks.

Tabel 23. Taotluste annotatsioonide sarnasusotsingute tulemused. Tabelis on taotluse ID, More Like This ning KNN tulemused.

Taotluse ID	More Like This	KNN
4c83b98e-e82d-45df-a40e-d0728f254efd	Keskkonna teema	Tervisega seotud tekstid
d1c1f19a-b8fc-4bb0-bd9e-5c20ae3a8f4c	Leidis üles stressi ja keskkonna	Leidis kohanemise teemad üles
748d99ce-2411-4b72-b181-a7a94b26bdc2	Ei leia	Vasted enam-vähem relevantssed
e978599f-cc16-47bc-b6ca-bc94fafdbe70	Tervise teema	Vasted enam-vähem relevantssed
a44e52fb-0280-4025-97d2-7eb5e9716f56	Ei leia	Ei leia
f4d18936-af9c-429c-9959-d24f87b577bc	Ei leia	Keeled ja kunstid
40922571-1a00-4067-ac02-84b5a39d50f0	Leidis ühe ehituse teksti	Leiab osa ehituse tekste
a657e8ce-91b0-458b-b56b-8d937e6ce7f7	Keskkonna teema	Tuvastas teema
cd8e4444-072e-46d9-a4d9-0e2430a0b882	Tuvastas keskkonna teema	Tuvastas teema

Nagu eespool tabelis kujutatud, siis kui kahte sarnasusotsingu varianti võrrelda, siis KNN sai paremaid tulemusi kui MLT (**tabel 23**). KNNi üheksast testannotatsioonist leiti neljale relevantssed segmendid, neli olid osaliselt sarnased ning üht ei leitud. MLT ei leidnud kolme sisendi puhul midagi relevantset ning ülejäänud sisendite puhul leidis samuti osaliselt sarnaselt segmente. Üks taotlus, mille puhul ei leitud ühtki vastet, oli hästi spetsiifilise teadusliku sisuga ning ilmselgelt polnudki ühegi andmekogumis oleva arengukavaga seotud.

6.3.2 Projektide annotatsioonide sarnasusotsingute tulemused

Analüüsi tegemiseks puudusid märgendatud testandmed. Andmete saamiseks leiti kümme projektide annotatsiooni, milles esines märksõna "arengukava" ning tehti eeldus, et need võiksid seostuda mõne segmendiga arengukavade andmestikust.

Tabel 24. Projektide annotatsioonide sarnasusotsingute tulemused. Tabelis on projekti ID, More Like This ning KNN tulemused.

Projekti ID	More Like This	KNN
e8249294-0a3d-4318-a061-1fbf2740f793	Esimene õige, teised mitte	Kõik õiged
35a4eb84-0c93-415b-bd63-521dd7a0e467	Ei leia	Enamik seostub teemaga
be7873d5-810d-4cfe-ba5d-27caaf6e7a22	Esimesed kolm õiged	Kõik õiged
42c016b4-908a-4854-8ca1-e098759d065b	Esimene õige, teised mitte	Enamik seostub teemaga
23c8b754-c0cd-4225-9399-5773bb1bce77	Teine õige	Kõik õiged
70c4df55-c2af-4bdd-9828-aa31a7dccf23	Enamik hariduse teemal	Kõik haridusega seotud
9c4a8692-f9e0-4c1a-9a6a-1bc7e2656d17	Viimane õige	Kõik julgeolekuga seotud
8bbc232f-0caa-4c1c-af12-9edaa440508f	Paar õiget	Esimene ja viimane kõige relevantsemad
3c4b7439-5173-489a-a371-f842e8b61c97	Üks õige keskel	Enamik seostub teemaga
49170795-de83-4254-9e1e-08c02f9963f1	Paar esimest õiged	Enamik seostub teemaga

Ka sel juhul on KNN otsing täpsem kui MLT (**tabel 24**). MLT leidis enamikel juhtudel osa relevantseid segmente, ühel korral ka ei leidnud üldse vastavaid segmente. KNN otsing seevastu andis pooltel juhtudel ainult relevantseid segmente või ülejäänud juhtudel oli enamik segmente relevantseid.

6.3.3 Sünteetiliste näidete sarnasusotsingute tulemused

Enamik sünteetilisi näiteid sai siduda spetsiifilise segmendiga, mis oli antud näite allikaks (välja arvatud teine näide, mis on hästi üldine). Seetõttu hinnati heaks need vasted, kus algne segment esines esimesena ning rahuldavaks need vasted, kus algne segment esines kuskil mujal vastetes või leiti ainult laiem teema.

Tabel 25. Sünteetiliste näidete sarnasusotsingute tulemused. Tabelis on näite kirjeldus, More Like This ning KNN tulemused.

Näite kirjeldus	More Like This	KNN
Sõnasõnaline kattuvus	Esimene leitud segment	Esimene leitud segment
Wikipedia-artikkel arengukavade kohta	Üldine jutt	Väga palju üldosi
Veidi muudetud sõnastusega ning laiendatud teemaga näide	Leiti ainult laiem teema	Esimene leitud segment
Tugevalt ümbersõnastatud	Ei leitud	Esimene leitud segment
Tugevalt ümbersõnastatud	Ei leitud	Ei leitud
Väga tugevalt ümbersõnastatud	Ei leitud	Ei leitud

Tulemustest nähtub, et mõlema sarnasusotsingute liigi puhul saaks otsingut kasutada nendel juhtudel, kui arengukava sisu sõnastust või lauseehitust on muudetud või täiendatud, aga mitte nii palju, et sisu oleks radikaalselt erinev. KNN leidis tihedamini täpse segmendi (**tabel 25**). Kahe allikast väga erineva keelekasutusega näite puhul ei leidnud kumbki otsing õiget allikat. Huvitav oli see, et esimese näite õige segmendi KNN sarnasusskoor oli 1,75 – üks näide projektide annotatsioonide hulgas oli sama sarnasusskooriga kui allikast kopeeritud näide. Teise, arengukavadesse mittepuutuva näite puhul leiti ikka sama arv väljundeid, mis sisaldasid hästi üldist teksti, näiteks sisukorda või eessõna.

6.4. Tulemuste kokkuvõte

Nii projektide annotatsioonid kui sünteetilised näited andsid piisava ülevaate arengukavadega kattumisest, kas siis spetsiifilise punktiga või temaatiliselt. Ka taotluste seas oli mõni annotatsioon, mis seostus arengukavades kirjeldatud teemadega. Mõlema sarnasusotsingu liigi puhul saaks otsingut kasutada nendel juhtudel, kui arengukava sisu sõnastust või lauseehitust on muudetud, aga mitte radikaalselt.

KNN vektorotsing suudab leida üles nii sisulise kui vormilise kattuvusega osad, isegi kui teemat on laiendatud. KNN andis kõigil testandmetel parema tulemuse kui MLT sarnasusotsing, tuvastades täpsemalt annotatsioonides või näidetes esinevad arengukavade teemad või otsesed viited arengukavade sisule. Seega oleks mõistlik seda otsinguliiki kasutada. Samas teeks isegi MLT kasutamine arengukavades vastavuste leidmise protsessi märksa lihtsamaks ja kiiremaks.

RavenDBs analüüsi kirjutamise hetkel KNN-otsingu võimalust pole, vektorite indekseerimisprobleemide tõttu, küll on aga palju teisi andmebaase, mis KNN-otsinguid toetavad ning rakendusi, mille abil oleks võimalik sarnasusotsingut andmebaasist olenemata sooritada.²⁷

6.5. Soovitused tulemuste parandamiseks

Arengukavadest vastavuse leidmise abilise loomisel oleks kasu sellest, kui arengukavade andmete töötlusel kasutada ainult sisukaid tekstiosid, ehk eemaldada viited, sisukorrad ja muud tekstid, mille olemasolu taotlustest ei otsita. Praegu sai osa lühematest tekstijuppidest välja sorteeritud, kuid hea oleks saada lahti kõigist tekstidest, millele pole vaja vastavust otsida. Põhjalikumate segmentide puhastust saaks teha regulaaravaldiste abil.

Kõik eksperimendis kasutatud tekstid olid eesti keeles, kuid vajadusel saaks sarnasusotsinguid teha ka inglise keeles. Selle eelduseks oleks ingliskeelsete arengukavade kasutuselevõtt ja samalaadse meetodi kasutamine. Inglise keele jaoks võib mõni muu vektormudel olla otstarbekam kui praegune kasutatud mudel.

Samuti oleks loodava abilise kvaliteedi jaoks vajalik see, et sisendiks kasutatavat taotlust ei esitata täispikkuses, vaid saaks kasutada sellest relevantsemaid osi/lõike, kuna sarnasusotsingute meetoditel on limiidid, kui suurt sisendit need suudavad otsida. Limiidi ületamisel näiteks KNN puhul ei anta ka veateadet, lihtsalt kasutatakse otsinguks ainult sisendi algusosa. Kui siiski oleks vaja kogu taotlust analüüsida, siis peaks selle eelnevalt segmentima (optimaalselt lehekülje pikkuseks või lõikudeks), seda saaks samuti teha regulaaravaldistega.

Selle analüüsi eksperimendid sooritati ühe lehekülje pikkuste segmentidega arengukavadest, kuid kindlasti oleks võimalik ka tagastada leitud vasted arengukavadest lõigetena, kui selle jaoks on vajadus. Selle jaoks oleks vajalik ära piiritleda, kuidas käituda loeteludega ja muude tekstiosadega, mis pole otseselt lõigu kujul. Lõikude kättesaamiseks saaks kasutada regulaaravaldisi. Kui segment oleks lõige kujul, võib kaaluda rohkemate metaandmete eraldamist arengukavadest ning segmentitud tekstiosade juurde metaandmete lisamist. Näiteks võib olla võimalik regulaaravaldiste abil tuvastada vahepealkirju, mille abil otsida ainult arengukavade sisulisest osadest, näiteks kokkuvõtetest. Sellega saaks andmetele lisada veel konteksti ning teha tagastatavad tekstilõigud lõppkasutajale arusaadavamaks.

Mõlemad sarnasusotsingu liigid leidsid iga sisendi vasteks kümme teksti arengukavadest olenemata sellest, kas esineb rohkem relevantseid arengukavade tekste või ei esine ühtki. Probleemi võiks proovida lahendada dünaamiliselt ning kasutada tagastavate vastete lävendina hoopis sarnasusskoori. Õige lävendi leidmine, millisest sarnasusskoori arvust on tagastatav vaste relevantne, nõuaks üsna palju eksperimenteerimist, seega ei saa selle osas täiendava analüüsita täpsemaid soovitusi pakkuda.

²⁷ Sarnasusotsingud sooritati ElasticSearch andmebaasis, kuid ka RavenDBs on *More Like This* vektorotsing, seega saaks Teadusagentuur kasutada juba tuttavas tööriistas sisseehitatud otsinguid, kasutades *Term Vectors* ja/või *More Like This* otsingut.

Teine variant vastete piiramiseks oleks kasutusele võtta mõni suur generatiivne keelemudel (*large language model*, edaspidi **LLM**) ja kasutada *retrieval augmented generation*-laadset lähenemist. Sisend ning saadud vasted saadetakse LLMile koos juhistega, et LLM valiks välja kõige õigema vaste. Arengukavade ja sisendi vastavuse puudumisel saaks LLM anda teada, et vastet ei leitud.

LLMi negatiivne pool on see, et pika sisendi puhul eelistab LLM pigem esimest või viimast otsingu vastet.²⁸ Seda probleemi praegu lahendatud veel pole. Teiseks kipuvad LLMid hallutsineerima (sel juhul näiteks olematust arengukavast vastet pakkuma), kuid seda saab vähendada, kui genereerimisparameetreid rangemaks muuta, näiteks temperatuuri langetada. Lisaks tuleb arvestada rahalise kuluga, mis oleneb sisend- ja väljundtekstide pikkusest ja spetsiifilise keelemudeli valikust. Näiteks kui sisend (kolm keskmise pikkusega lehekülge arengukavadest, päring ning võrreldav lõik taotlusest) on 1242 sõne pikkune, siis võiks ühe päringu sisendi kulu olla 0,11–2,3 senti. Kui eeldada, et väljundi pikkus on umbes 300 tokenit, siis väljundi maksumus jääks vahemikku 0,04–2,07 senti. Samuti võivad tekkida juriidilised küsimused seoses taotluste konfidentsiaalsusega.

Tabel 26. Ülesanded, mida võiks tulevikus teostada, koos ülesande mahuhinnanguga.

Ülesande kirjeldus	Eeldatav maht
Arengukavade segmentide puhastamine	30 tundi
Arengukavade lõikudeks segmentimine	50 tundi
Vahepealkirjade tuvastamine	40 tundi
Tagastavate dokumentide arvu sidumine sarnasusskoori lävendiga	30 tundi

²⁸ Liu, Nelson F. jt 2023. Lost in the Middle: How Language Models Use Long Contexts. arXiv:2307.03172

7. Publikatsioonist populaarteadusliku kokkuvõtte tegemine

Uurisime publikatsioonist automaatsete kokkuvõtete tegemist erinevate LLM-mudelite abil. LLM-mudeld kasutavad kokkuvõtte tegemisel lihtsamat keelt ning vajadusel lisavad selgitavaid andmeid, alternatiivseid meetodeid LLM-i kasutusele praegu pole. Lähteülesandest tulenevalt tuli lahendust testida nii annotatsioonide kui ka publikatsioonide täistekstide peal.

7.1. Alusandmed

Valdavalt on eksperiment üles ehitatud annotatsioonidele, sest laiem täispikkuses publikatsioonide analüüs oleks hästi suure analüüsimahuga. Testimiseks kasutatakse kolme täispikkuses publikatsiooni eesti keeles, mis on kõik keeleteaduse, täpsemalt keeletehnoloogia valdkonnast. Samuti kasutatakse nelja täispikkuses publikatsiooni inglise keeles, mille autorid on ETAGist. Et kindlustada publikatsioonide sisukokkuvõtete faktipõhisus, on läbi loetud nii täispikkuses publikatsioonid kui sisendiks olevad annotatsioonid. Publikatsioonide annotatsioonide näideteks kasutatakse OpenAlexist saadud publikatsioonide puhastatud annotatsioone, täpsemalt 16 eestikeelset annotatsiooni, mis kuuluvad OpenAlexi järgi erinevatesse valdkondadesse. Lähemal uurimisel osutus enamik tekste hoopis keeleteaduse valdkonda kuuluvateks. Kasutatud dokumentide DOI/ETISE lingid on [lisas 4](#) ning dokumentide arv ning pikkus on kirjeldatud alljärgnevas tabelis (**tabel 27**).

Tabel 27. Populaarteaduslike publikatsioonide alusandmete dokumentide tüüp, arv ning sõnede summa.

Dokumendi tüüp	Dokumentide arv	Sõnede summa kokku	Tähemärkide summa kokku
Täispikkuses publikatsioon (artikkel)	7	52 595	355 823
Publikatsiooni annotatsioon	16	24 821	157 611
Kõik dokumendid (summa)	23	77 416	513 434

7.2. Meetodid ja mudelite valik

LLM eri mudelite kasutus on ülesande püstituse lahendamiseks ainus valik, kuna muud meetodid pühenduvad ainult kokkuvõtete tegemisele, mitte sõnavara lihtsustamisele ning seletamisele.

Mudelid, mille sooritust hinnatakse ja omavahel võrreldakse, on:

- OpenAI GPT-4-turbo²⁹ (edaspidi GPT 4T)
- OpenAI GPT-3.5-turbo³⁰ (edaspidi GPT 3.5T)
- Anthropic Claude 3.5 Sonnet³¹ (edaspidi Sonnet)
- Anthropic Claude 3 Opus³² (edaspidi Opus)
- Google Gemini Pro 1.5³³ (edaspidi Gemini)

Mudelid on välja valitud eelnevate analüüside tulemusel, kus lasti erinevatel LLM-mudelitel teha kokkuvõtteid ajakirjanduslikest tekstidest. Kuigi ajakirjanduslik žanr on teaduslikest tekstidest erinev, siis eelnevad tulemused annavad siiski üldiselt arusaama eestikeelsete kokkuvõtete tegemise oskusest. ETAGi autorite täispikkuses publikatsioonide põhjal kokkuvõtete tegemiseks kasutatakse ainult GPT 4T, Sonnet ja Gemini mudeleid.

Kõigile mudelitele edastatakse samad päringud API kaudu. Hindamise käigus vaadeldakse genereeritud kokkuvõtete kvaliteeti ning juhendistes kinnipidamist. Samuti võrreldakse arvestatavat hinda. Lisaks arvestatakse võimalike sisendi või väljundite limiitide ning tõrgetega genereerimisel. Samuti tuuakse välja ohtlikud kohad, kus mudel hakkab hallutsineerima (ise infot lisama või muutma).

7.2.1 Päringud

Kasutati kahte tüüpi päringuid:

Esiteks lihtne päring probleemipüstitusega:

*Oled spetsialist, kes kirjutab teaduspublikatsioonide sisukokkuvõtteid eesti keeles.
Kokkuvõtte sisu peaks olema tavainimesele arusaadav, see ei tohiks sisaldada žargooni.
Kõik keerulisemad väljendid võiks olla seletatud lihtsas keeles.
Kokkuvõtte pikkus: 300–400 sõna.*

Teiseks päring, kus on samad juhendid nagu esimeses näites, kuid lisaks veel üks näide annotatsioonist ning kokkuvõttest. Näiteks valiti võimalikult värske Keele ja Kirjanduse artikli annotatsioon, et see puuduks vaadeldavate annotatsioonide hulgast. Näite tekitamiseks kasutati GPT mudeli abi, kuid lühikese pikkuse (257 sõna) ja puuduva sisu tõttu oli vaja näide käsitsi teha korrektseks, et see vastaks päringu tingimustele (vastaks õigele pikkusele ning kataks puuduva sisu). Päringus kasutatud näiteannotatsioon ning -kokkuvõte on lisas 5. Enne ETAGi töötajate

²⁹ <https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo> - GPT-4-turbo, keelemudel

³⁰ <https://platform.openai.com/docs/models/gpt-3-5-turbo> - GPT-3.5-turbo, keelemudel

³¹ <https://www.anthropic.com/news/claude-3-5-sonnet> - Claude 3.5 Sonnet, keelemudel

³² <https://www.anthropic.com/news/claude-3-family> - Claude 3 Opus, keelemudel

³³ <https://deepmind.google/technologies/gemini/pro/> - Gemini Pro 1.5, keelemudel

publikatsioonide põhjal kokkuvõtete tegemist parandati näitekokkuvõttes algselt sees olnud kirjaviga, kuid see võis mõjutada teisi genereeritud kokkuvõtteid.

7.2.2 Hinna arvestamine

Kõik mudelid on hinnastatud tokeni-põhiselt. LLMide valdkonnas tähendab token üht sõne või eesti keele puhul tihti ka sõnest väiksemat ühikut. OpenAI mudelite puhul kasutati sisendi ja väljundi tokenite kättesaamiseks *tiktoken*³⁴ Python teeki. Nii Anthropic kui Google mudelitel olid sisseehitatud lahendused tokenite loendamiseks, seega tokenite arvu sai päringu tegemisel salvestada.

7.3. Tulemused

7.3.1 Genereeritud sisukokkuvõtete kvaliteet

Enamik sisukokkuvõtteid olid arusaadavas eesti keeles, osadel juhtudel esines kohmakust, veidraid keelendeid või kirjavigu. Enamikel mudelitel esines ka vähemalt paar korda kokkuvõte, mida hinnati halvaks (kohmakuse, veidrate keelendite ja kirjavigade koosmõjul). Ladusama keelega kokkuvõtted olid Opus ja Sonnet mudelitel, neile järgnes GPT 4T. Kõige kohmakamate kokkuvõtetega olid Gemini ja GPT 3.5T. Kvaliteedi tulemused ning kokkuvõtted on lisatud eraldi tabelisse³⁵.

Kokkuvõtted annotatsiooni põhjal kippusid annotatsiooni teistes sõnades ümber jutustama. Täismahus publikatsioonide puhul pidi mudel valima, mis on kõige relevantsem informatsioon, üldjuhul oli kokkuvõte täismahus publikatsioonide lõpuosadest (arutelu, kokkuvõte jms). Polnud olulist kvaliteedierinevust täismahus ning annotatsioonide põhjal tehtud kokkuvõtete vahel. Genereeritud kokkuvõtete põhjal eristusid teiste seast Gemini keelemudel, mis genereeris rohkem näiteid erinevaid keeruliste sõnade kohta, ning Opus ja Sonnet keelemudelid, mis genereerisid rohkem struktureeritud kokkuvõtteid, kus sisu oli edasi antud lihtsasti hõlmatava nimekirjana.

Žargon ning erialaspetsiifiline sõnavara oli mudelite poolt üldjuhul tuvastatav. Enamikel juhtudel asendasid mudelid ära vaid ühe või paar oskussõna- või väljendit, mõnel juhul genereeris mudel kokkuvõtte teksti lihtsustamata. Ilmselt kui mudelite parameetreid muuta või esitada konkreetne palve teatud termini selgitamiseks või asendamiseks, siis saaks selle lahendada. Näitega päringu puhul võis mõjutada tulemusi üks kirjaviga, mis näitesse sisse jäi, seetõttu olid ka selle tulemused kehvemad.

Mudeleid hinnati keele ladususe ning kokkuvõtte lihtsuse põhjal süsteemis, kus ühe päringu eest võis saada maksimaalselt ühe punkti (kokku 19 punkti). Eraldi hinnati ETAGi töötajate publikatsioone (**tabel 29**). Hallutsineerimise puhul sai mudel genereeritud sisu eest 0 punkti, võimalik oli saada ka 0,75 punkti, kui esinesid väiksemad kirjavead/ladususevead või 0,5 punkti, kui muidu oli kokkuvõte ladus, kuid kokkuvõte polnud oluliselt lihtsam. Enamik genereeritud

³⁴ <https://github.com/openai/tiktoken> - tiktoken, tokeniseerija

³⁵ Keelemudelite koostatud kokkuvõtete hindamistabel.xlsx

kokkuvõtetest said kas 0,5 või 0,75 punkti. Kvaliteedi põhjal oleks kolm parimat mudelit Sonnet, Gemini ja GPT4 (**tabel 28**). Erinevate kasutatud päringute puhul pole kvaliteedis suurt vahet, Gemini ja GPT 4T puhul näite kaasamine päringusse oli korrelatsiooniliselt halvema kvaliteediga kokkuvõtetele.

Tabel 28. Parimad mudelid keele ladususe ja kokkuvõtte lihtsuse põhjal.

Mudeli nimi	Päring näiteta	Päring näitega
Claude 3.5 Sonnet	10	10,25
<i>Gemini 1.5 Pro</i>	<i>10,5</i>	<i>8,75</i>
GPT-4 Turbo	10	8,25

ETAGi töötajad hindasid omaenda artiklite kokkuvõtteid, sisendiks antud artikleid oli neli ning artiklid olid kogupikkuses esitatud. Kokkuvõtete genereerimiseks kasutati vaid eelnevas analüüsis selgunud parimaid mudeleid ehk samu, mis välja toodud **tabelis 28**. ETAGi teadlaste hinnangul olid parimad mudelid Gemini 1.5 Pro ning Claude 3.5 Sonnet. Maksimaalsest kaheksast punktist (kahe päringu peale kokku) sai Gemini 5,25 ja Sonnet 5 punkti. GPT 4 hinnati üldiselt halvemaks (**tabel 29**). ETAGi töötajate hindamistabel erineb eelnevast tabelist selle põhjal, et hallutsineerimise ega vale keele kasutamise eest pole automaatselt antud 0 punkti.

Tabel 29. Parimad mudelid keele ladususe ja kokkuvõtte lihtsuse põhjal, ETAGi töötajate hinnang oma publikatsioonide kokkuvõtetele.

Mudeli nimi	Päring näiteta	Päring näitega
<i>Claude 3.5 Sonnet</i>	<i>2,75</i>	<i>2,25</i>
Gemini 1.5 Pro	2,75	2,5
GPT-4 Turbo	0,75	1,5

Kokkuvõtteid hindavate ETAGi töötajate poolt tuli tagasiside, et kuigi genereeritud kokkuvõtete sõnastus oli konarlik, siis tundusid need siiski mitte-ekspertidele lugemiseks lihtsamad ning ka hallutsineerimist esines üsna vähe. Näitekokkuvõtete päringu puhul oli näha, et kokkuvõtete kvaliteet muutus: GPT-4 mudeli tulemused läksid veidi paremaks, kuid teistel üldiselt langesid. Kuigi näiteta päringul oli kokkuvõtete kvaliteet parem, siis genereeriti ETAGi hinnangul näitega päringu puhul kaks kõige kõrgemini hinnatud kokkuvõtet.

7.3.2 Juhenditest kinnipidamine

Kõik mudelid eksivad genereeritud sõnade arvuga, genereerides lühemaid kokkuvõtteid kui palutud. Tavakasutajal oleks võimalik teha järgmine päring tagasisidena, et kokkuvõte oleks pikem, kuid API kasutamise tõttu esitame üksikpäringud ilma ajaloota. Praegusel ajal on ka teaduslik konsensus, et LLM mudelid ei saa täpsemate pikkusnõuetega päringutega hakkama ning otsest lahendust sellele pole.

Näitega päringu puhul oli kokkuvõtte keskmine sõnade arv 4–42 sõna enam kui näiteta päringu puhul (**tabel 30**). Suurim vahe oli Opus ning väiksem GPT 4T mudelil. Keskmine sõnade arv on kõige suurem GPT 4T mudelil ja kõige väiksem GPT 3.5T mudelil.

Tabel 30. Genereeritud kokkuvõtete keskmine ja maksimaalne sõnade arv mudeli järgi.

Mudeli nimi	GPT-4 Turbo*	GPT-3.5 Turbo	Claude 3.5 Sonnet*	Claude 3 Opus	Gemini 1.5 Pro*
Keskmine sõnade arv	205,2	168,0	175,7	174,7	176,7
Maksimaalne sõnade arv	253	246	258	234	272

Sonnet mudel eksis ka pidevalt kokkuvõtete keelenõudega, genereerides täispikkade publikatsioonide jaoks hoopis ingliskeelseid kokkuvõtteid. Sellegi probleemi saaks lõppkasutaja lahendada lihtsalt tagasisidet andes.

7.3.3 Mudelite hind

Kõik mudelid on hinnastatud *tokeni*-põhiselt. LLM-ide valdkonnas tähendab *token* üht sõne või eesti keele puhul tihti ka sõnest väiksemat ühikut. Hinnaandmed on lehel Artificial Analysis³⁶, mis kogub statistikat erinevate suurte keelemudelite kohta.

Tabel 31. Mudelite hinnad ühe miljoni tokeni kohta (dollarites).

Mudeli nimetus	Sisendi hind ühe miljoni <i>tokeni</i> kohta (dollarites)	Väljundi hind ühe miljoni <i>tokeni</i> kohta (dollarites)
GPT-4 Turbo	10,00	30,00
GPT-3.5 Turbo	0,50	1,50
Claude 3.5 Sonnet	3,00	15,00
Claude 3 Opus	15,00	75,00
Gemini 1.5 Pro	3,50	10,50

Kõige kallimad mudelid on Opus ning GPT 4T (**tabel 31**). Kõige soodsam on GPT 3.5T. Mudelite hinnad varieeruvad üsna palju, samuti on sisend*tokenid* (päringu ja juhiste tekst) ja väljund*tokenid* (genereeritud sõnumid) eri hinnaga, tavaliselt on väljund*tokenid* kallimad.

³⁶ <https://artificialanalysis.ai/models> - Artificial Analysis, suurte keelemudelite võrdluskeskond

* Antud mudelite puhul on arvestatud ka hiljem lisatud neli täispikkuses dokumenti.

Järgnevas tabelis on ülevaade eksperimendi käigus saadetud ning saadud *tokenitest* ning umbkaudsest maksumusest (**tabel 32**). Enamik sisendeid sai mudelitele kaks korda ette anda (välja arvatud pikemad täispublikatsioonid pikkuse limiitidega mudelitel), samuti genereeriti kõigi sisendite kohta kaks kokkuvõtet. Maksumus on arvatud **tabelis 31** toodud hindade ning *tokenite* arvu põhjal, valuuta on ümber arvestatud hetkekursi järgi (1 USD = 0,92 EUR).

Nagu näha, on isegi samade sisendite põhjal *tokenite* arv erinev. See tuleneb sellest, et mudelitel on kasutusel erinevad *tokenisaatorid* ehk tööriistad, mis teksti *tokeniteks* protsessivad. Näiteks OpenAI *cl100k_base* *tokeniseerija* järgi (kasutusel GPT 4T ja GPT 3.5T mudelitel) oleks sõnum "Tere maailm!" kuue *tokeni* pikkune: "T", "ere", "ma", "ail", "m" ja "!".

Tabel 32 näitab, et Sonnet ja eriti Opus mudelitel on sisendi *tokenite* kogus märkimisväärselt väiksem kui ülejäänud LLM-mudelitel. Ka väljund*tokenites* on erinevus, nimelt nii GPT 3.5T kui Gemini³⁷ on genereerinud vähem *tokeneid*, samas moodustasid need kaks mudelit keskmiselt kõige lühemad kokkuvõtted. Kõige rohkem väljund*tokeneid* on GPT 4T mudelil, millel olid ka keskmiselt kõige pikemad kokkuvõtted.

Tabel 32. Eksperimendi jooksul kasutatud *tokenite* arv ning umbkaudne hind eurodes.

	Sisend- <i>tokenid</i>	Väljund- <i>tokenid</i>	<i>Tokenid</i> kokku	Sisendi hind (€)	Väljundi hind (€)	Hind kokku (€)
GPT-4 Turbo*	324 260	30 833	355 093	2,98	0,85	3,83
GPT-3.5 Turbo	179 384	19 006	198 390	0,08	0,03	0,11
Claude 3.5 Sonnet*	287 313	23 623	310 936	0,79	0,33	1,12
Claude 3 Opus	105 204	20 130	125 334	1,45	1,39	2,84
Gemini 1.5 Pro*	323 391	22 506	345 897	1,04	0,21	1,25

Kui lähtuda vaid soodsast hinnast, siis oleks parim variant kasutada GPT 3.5T mudelit, veidi halvemad on Gemini ning Sonnet ning kõige halvemad GPT 4T ja Opus.

³⁷ Gemini mudelit tuleks võrrelda tabelis teiste tärnidega märgitud mudelitega, millel oli rohkem sisendeid.

* Antud mudelite puhul on arvestatud ka hiljem lisatud neli täispikkuses dokumenti.

7.3.4 Limiidid, tõrked, hallutsineerimine ning muud märkused

Mudelitele kehtestatud sisendi pikkuse limiidid kippusid kokkuvõtete genereerimist takistama ainult täispikkade publikatsioonide analüüsil, see mõjutas Sonnet, Opus ning GPT 3.5T mudeleid. Samas oli üks publikatsioon ka hästi pikk (üle 100 lk) ning nii GPT 4T kui Gemini said sellest kokkuvõtte tegemisega hakkama.

Geminil oli ühes minutis esitatavate päringute limiit rangem kui teistel mudelitel, seetõttu pandi Gemini päringutele vahele 60-sekundilised pausid, ka Sonnetil esines sarnane situatsioon. Korra esines Geminil tõrge, kus genereeriti tühi vastus, tõenäoliselt jäi genereeritud vastus turvafiltrisse kinni. Uuel genereerimisel sama tõrget ei tekkinud.

Üsna palju esines kokkuvõtete genereerimisel hallutsinatsiooni. Ühel korral lisas Opus ühele artiklile vale autori nime (*Keeleõpetaja Piret Põldver uurib oma doktoritöös [...]*), Opus kasutas ka viidet "*Sinu kirjeldatud artikkel*" ühe kokkuvõtte alguses ning kirjeldas samas kokkuvõttes ülesannet: "*Teaduspublikatsioonid võivad tihti olla tavainimesele raskesti mõistetavad, kuna sisaldavad erialast terminoloogiat.*"

Gemini mudel mõtles välja segaseid näiteid modaaltegusõnade kohta: "*Saama*" ja "*võima*" on *modaalverbide ehk abisõnade tüüp, mis väljendavad tegevuse võimalikkust, lubatavust või vajadust. Näiteks võime öelda "Ma saan süüa" (võimalus) või "Ma ei saa süüa" (lubamatu) või "Sa pead saama süüa" (vajadus). Samuti võime öelda "Ma võin süüa" (lubamatu).* Ka GPT 4T mudel eksis korra, asendades Matsalu lahe roostiku-uuringus *tsükloni* sõnaga *tsunami*.

Sisuliselt kõik mudelid välja arvatud GPT 3.5T eksisid annotatsiooni puhul, kus oli nimekiri koodivahetuslikest strateegiatest, kuid strateegiad olid kirjeldamata. Ühe strateegia puhul eksisid mudelid selle kirjeldamisel, sest täispikkuses publikatsioonis oli antud strateegia kirjeldus hoopis teistsugune, ehk siis mudelid mõtlesid sellele strateegiale seletuse ise välja. Publikatsioonide vigade parandamiseks oleks vaja toimetajat, kes on kas antud teemaga rohkem kursis või täispublikatsiooni sirvinud.

7.4 Tulemuste kokkuvõte

Kui sisendiks anda ainult publikatsiooni annotatsioon, siis saavad kõik mudelid ülesandega vähemalt osaliselt hakkama, enamik moodustatud kokkuvõtetest on üsna ladiusad ning veidi lihtsama keelega. Kui sisendiks on täispikkuses publikatsioon, siis võivad segada mudelite limiidid (näiteks GPT 3.5T, Opus või Sonnet puhul), mille tõttu jääb kokkuvõtte genereerimata.

Tulemuste põhjal nähtub, et kõige rohkem sobiksid antud ülesande lahendamiseks Sonnet või Gemini mudel. Mõlemal on omad plussid ja miinused. Mõlemad on samas hinnaklassis: mitte kõige kallimad, kuid mitte kõige soodsamad, ning mõlemad paistsid kvaliteedi hindamisel silma. Sonnet tekitas hästi struktureeritud kokkuvõtteid ning Gemini kasutas keerulisemate mõistete puhul hästi palju näiteid. Mõlemal on kalduvus hallutsineerida ning Gemini puhul esines ka korra tõrge

genereeritud sõnumi filtreerimise tõttu (kuigi Gemini mudeli puhul peaks saama filtreerimisparameetreid täpsustada). Gemini saab samas hakkama ka hästi pikkade sisenditega, mille puhul Sonnet läheb üle inglise keelele või saadab veateate.

Kui sobiks ka veidi kallim keelemudel, siis ka GPT 4T genereeritud tekst oli kvaliteetne ning sisendi pikkus samuti suur, kuid samamoodi nagu teised on GPT 4T hallutsineerimiskalduvustega. Kui oluline on vaid hind ning hallutsineerimiskindlus, siis GPT 3.5T oleks parim valik, kuid tuleb arvestada, et selle mudeli sisendi limiit on üsna lühike, see genereerib pigem lühemaid tekste ning ka keelekasutust tuleks siis rohkem korrigeerida.

7.5 Soovitused tulemuste parandamiseks

Tulemusi saaks kindlasti parandada, kui muuta mudelite genereerimisparameetreid (näiteks temperatuuri, mis kontrollib loovust), selle kaudu saaks näiteks vähendada hallutsineerimist. Kui mõni termin jääb mudeli poolt lihtsustamata või seletamata, siis saaks probleemi järgmises vestlusvoorus lahendada. Kindlasti saaks katsetada ka teiste päringutega, mis on optimeeritud soovitud kokkuvõtte saamiseks. Päringule saaks lisada näiteks lause, et vähendada hallutsineerimist keeruliste väljendite seletamisel. Et kokkuvõtte oleks faktipõhine ning ladus, peaks LLMi moodustatud kokkuvõtted üle vaatama keeletoimetaja või -korrektor.

Kui praegu jäävad tulemused siiski liiga halvaks, siis õnneks on LLM kiirelt arenev valdkond ning uusi mudeleid avaldatakse sisuliselt iga paari kuu järel. Samas on risk, et ka hinnastus või valuutakurss võivad muutuda, mis võib mudeli kasutamise teha ebamõistlikuks. On lootus, et tekivad ka vabavaralised mudelid, mis tulevad eesti keelega paremini toime (näiteks eestikeelse LLM mudeli Llammas³⁸ edasiarendus või Turu ülikoolis loodav Euroopa keelte LLM³⁹).

Kindlasti tuleb arvestada ka juriidiliste küsimustega, kuigi analüüsi kirjutamise ajal eksperimendis käsitletud mudeleid haldavad ettevõtted ei kogu äriklientide sisendandmeid edasise treenimise eesmärgil, siis võib see muutuda.

Kui tahta sisukokkuvõtet PDFi põhjalt, mitte tekstilise sisendi alusel, siis kõik kasutatud mudelid peaks seda võimalust toetama, kuid on limiidid sisenddokumendi mahule. OpenAI-l saab olla sisendiks PDF kuni 512 MB, Google'il kuni 100 MB, Anthropicul kuni 30 MB.

³⁸ [Kuulmets, Hele-Andra jt 2024. Teaching Llama a New Language Through Cross-Lingual Knowledge Transfer. Findings of the Association for Computational Linguistics: NAACL 2024, lk 3309-3325.](#)

³⁹

<https://www.utu.fi/en/news/press-release/university-of-turku-and-silogen-launch-consortium-to-build-the-worlds-largest> - Turu Ülikool, loodava LLM-i kohta.

Tabel 33. Ülesanded, mida võiks tulevikus teostada, koos ülesande mahuhinnanguga.

Ülesande kirjeldus	Eeldatav maht
Mudelite parameetritega eksperimenteerimine	40 tundi
Uute päringutega katsetamine (hallutsineerimise vähendamiseks)	10–30 tundi (olenevalt päringute arvust)

8. Eesti Teadusinfosüsteemi tekstiotsingute täiendamine

Eesti Teadusinfosüsteemi (ETISE) tekstiotsingud ei arvesta käändelõppudega, samuti ei leita sünonüüme. Tekstiotsingute täiendamisest oleks kasu nii ETAGi töötajatele kui ka näiteks ministeeriumitele otsuste tegemiseks. Allpool on välja toodud võimalused otsingute parandamiseks.

8.1 Võimalikud meetodid probleemi lahendamiseks

Teadusinfosüsteemi (ETISE) tekstiotsingut on võimalik täiendada ja parandada mitmel erineval viisil. Üheks viisiks oleks kasutada sarnasusotsinguid. Sarnasusotsingud leiavad üles nii sünonüüme kui ka käänatud sõnu või alternatiivseid fraase. Sarnasusotsingud on pikemalt kirjeldatud arengukavadest vastavuste leidmise abilise analüüsis. Vektorotsingu jaoks oleks vaja otsitavad andmed vektoriseerida vektoriseerimismudeli abil, nagu näiteks BGE-M3. Otsinguid saaks sooritada vektoriseeritud välja põhjal. Sarnasusotsingud väljastavad alati kas ühe kindla arvu vasteid või siis kõik vasted, mis ületavad teatud sarnasusskoori. Enne sarnasusotsingute implementeerimist tuleks uurida, kumb vastete väljastamise viis kasutajatele rohkem sobiks.

Lisaks sarnasusotsingutele võib kaaluda ka *retrieval augmented generation* põhimõttel mõne LLM-mudeli kasutamist, mis valiks saadud vastete hulgast välja kõige täpsema. Selle lähenemisega seostuvad aga olulised kulud, kui LLM-põhist vektorotsingut kõigile kasutajatele võimaldada.

Kolmas võimalik lähenemine oleks otsitava välja (või väljade) ja otsisõnade automaatne lemmatiseerimine, kasutades näiteks *Texta Multilingual Processorit* (MLP)⁴⁰ või *estNLTK*⁴¹ teeki eesti keele jaoks. Sellisel juhul saaks kasutajad otsida näiteks sõna *karudega* ning vastusena kuvatakse kõik artiklid, kust leiti *karu*.

Teadusagentuuri spetsialistidele võib otsingu mugavamaks teha ka otsingusõnale või -fraasile sisult sarnaste sõnade soovitamise *Texta Toolkitis Lexicon Miner*⁴² kaudu. *Lexicon Miner* kasutaks sisendiks Teadusagentuuri andmetel treenitud sõnavektoreid.

⁴⁰ <https://pypi.org/project/texta-mlp/> - Texta Multilingual Processor, Pythoni teek keeleandmete töötlemiseks

⁴¹ <https://github.com/estnltk/estnltk> - EstNLTK, Pythoni teek keeleandmete töötlemiseks

⁴² https://docs.texta.ee/lexicon_miner.html - Texta Lexicon Miner, tööriist sarnaste sõnade leidmiseks

Tabel 34. Ülesanded, mida võiks tulevikus teostada, koos ülesande mahuhinnanguga.

Ülesande kirjeldus	Eeldatav maht
Detailsema sarnasusotsingute analüüsi läbiviimine	10–40 tundi (oleneb andmete mahust)
Lemmatiseerimise katse läbiviimine	10–40 tundi (oleneb andmete mahust)

9. Taotlustele sobivate hindajate leidmine ja huvide konfliktiga hindaja välistamine

Kuna taotlustele korrektsete valdkondade määramine osutus keeruliseks probleemiks, siis selles analüüsis ei teostata täiesmahus taotlustele sobivate hindajate leidmist ega huvide konfliktiga hindaja välistamist. OpenAlexist saab pärida publikatsioonide kirjeid kasutades valdkonna ID-d või autori nime/ID-d, kuid pikem töövoog kirjeldus ning saadud andmete kvaliteedi hindamine jääb tegemata.

Oluliseimad probleemid võivad tekkida autori nime ühestamisel. Autoritel võivad olla valed aliased ning ühel autoril võib olla mitu ID-d, kuid tuleks leida ainult need isikud ja ID-d, kellel või millel on seos menetluses taotlusega. Näiteks on Tartu Ülikooli õppejõul Mark Fišelil OpenAlexis kaks ID-d nimega *Mark Fišel*, samuti ID nimega *Mark Fishel* (erineva ID-ga on nimekaim Mark A. Fishel Ameerikas) ning viimase IDga on seotud aliased *M. Fishel*, *Mark Fishel* ja *Ondřej Bojar*.

9.1 Annotatsiooni ja valdkonna põhjal sarnaste publikatsioonide leidmise katse

Taotlusele hindajate leidmise eelduseks oleks määrata taotlusele valdkond, mille alusel OpenAlexist andmed pärida. Kuna ühes valdkonnas võib käsitletavaid teemasid ning publikatsioone olla suur hulk, siis tuleks tuvastada taotlusele kõige sarnasemad publikatsioonid. Selleks saaks kasutada kas märksõna- või KNN-otsingut, viimast ka testiti kogutud parasitoloogia valdkonna publikatsioonide annotatsioonidega.

Alusandmetena kasutati peatükis 5.4.1.2 kogutud OpenAlexi andmeid, mis vektoriseeriti samamoodi nagu peatükis 6.2.1 kirjeldatud arengukavade segmendid. Kokku oli vektoriseeritud dokumente 9354, vektoriseeritud välja sisuks sai ingliskeelne annotatsioon. Vektoriseerimiseks kasutati ühte GPU-d, sarnase riistvarata on protsess aeglasem või mahukama andmestiku puhul sisuliselt võimatu. Sooritati KNN sarnasusotsing, mille sisendiks oli ETISE [projekti](#) pealkiri ning annotatsioon (mõlemad inglise keeles).

Kokku leiti viis kõige sarnasemat dokumenti sisendile, tulemused leiab [lisast 6](#). Sisend uuris lõhilaste (kalade) parasiite seoses kliimasoojenemisega, keskendudes teatud parasiidi ning selle põhjustatud haiguste tuvastamisele ning parasiidi ökoloogia uurimisele.

Kõik KNN-otsinguga leitud sarnased annotatsioonid käsitlevad parasiite ning kliima soojenemist, kuigi täpsemalt kaladega seotud fookust leitud publikatsioonides ei esinenud. Enamikes annotatsioonides on ka erladi välja toodud veekogud. Kalu on mainitud parasitoloogia andmestikus 120 annotatsioonis ning lõhilasi 19 annotatsioonis, kuigi neis on fookus teistel parasiitidel ning

parasiithaigustel. Projektis mainitud parasiiti parasitoloogia andmestikus ei esine. Seetõttu on projekti kirjelduse vormist tulenevalt leitud veidi üldisemad publikatsioonide annotatsioonid, selle asemel, et anda vasted mõne teise lõhilaste parasiidi kohta. Otsingutulemused näivad hindajate leidmiseks sobilikud. Kui dokumendid on juba kogutud, siis saaks vajadusel kasutada ka märksõnaotsinguid, kui leitud vasted ei tundu piisavalt spetsiifilised.

Edasises töövoos võiks eraldada leitud publikatsioonide andmete seast autorite nimed, ning teha kontroll, kas leitud dokumentide autorid on taotluse esitanud autoritega kunagi teaduslikes publikatsioonides kaastööd teinud. Selleks saaks kasutada näiteks OpenAlex API päringuid autori/autoritega seotud publikatsioonide kohta ning kontrollides tagastatud publikatsioonide autorite nimekirju.

Tabel 35. Ülesanded, mida võiks tulevikus teostada, koos ülesande mahuhinnanguga.

Ülesande kirjeldus	Eeldatav maht
Taotlustele sobivate hindajate leidmise analüüs	25 tundi
Andmete hõivamine OpenAlexist (miljon dokumenti) ⁴³	60 tundi
OpenAlexi andmete töötlemine ja vektoriseerimine sarnasusotsingus kasutamiseks (miljon dokumenti)	60 tundi
Huvide konfliktiga hindaja välistamine analüüs	25 tundi

⁴³ Arvestatud on andmete hõivamine API kaudu, maksimaalselt saab teha 100 000 päringut päevas, ühes päringus võib pärida nt 200 kirjet. Suurema koguse andmeid saab hõivata ka OpenAlex kuvatõmmist kasutades.

10. Soovituslikud ülesanded ja mahuhinnangud analüüsi põhjal

Tabel 36. Soovituslikud ülesanded ja mahuhinnangud analüüsi põhjal.

Algne analüüs	Andmestik	Ülesande kirjeldus	Eeldatav maht
Taotlustele valdkondade määramine	OpenAlex / ETIS projektide ingliskeelsed annotatsioonid	Filtreerida välja ainult metaandmeid sisaldavad sisukokkuvõtted regulaaravaldistega	30 tundi
Taotlustele valdkondade määramine	OpenAlex / ETIS projektide ingliskeelsed annotatsioonid	Koodi või sümboleid sisaldavate kirjete puhastamine regulaaravaldistega	40 tundi
Taotlustele valdkondade määramine	OpenAlex	Lisaandmete hõivamine	40–60 tundi (olenevalt andmemahust)
Taotlustele valdkondade määramine	OpenAlex	Valdkondade käsitsi märgendamine	150–200 tundi (olenevalt andmemahust)
Taotlustele valdkondade määramine	ETIS projektide ingliskeelsed annotatsioonid	Andmete töötlemine	15 tundi
Taotlustele valdkondade määramine	ETIS projektide ingliskeelsed annotatsioonid	Tagger Mudelite treenimine ja valideerimine	30 tundi
Taotlustele valdkondade määramine	OpenAlex / ETIS projektide ingliskeelsed annotatsioonid	Valdkondade tunnuste korrigeerimine (stoppsõnade lisamine Tagger Group mudelitele)	30–45 tundi (olenevalt andmemahust)
Arengukavade vastavusanalüüsi abiline	Arengukavad	Arengukavade segmentide puhastamine	30 tundi
Arengukavade vastavusanalüüsi abiline	Arengukavad	Arengukavade lõikudeks segmentimine	50 tundi
Arengukavade vastavusanalüüsi abiline	Arengukavad	Vahepealkirjade tuvastamine	40 tundi

Arengukavade vastavusanalüüsi abiline	Arengukavad	Tagastavate dokumentide arvu sidumine sarnasusskoori lävendiga	30 tundi
ETIS otsingute täiendamine	-	Detailsema sarnasusotsingute analüüsi läbiviimine	10-40 tundi (oleneb andmete mahust)
ETIS otsingute täiendamine	-	Lemmatiseerimise katse läbiviimine	10-40 tundi (oleneb andmete mahust)
Populaarteaduslike kokkuvõtete genereerimine	-	Mudelite parameetritega eksperimenteerimine	40 tundi
Populaarteaduslike kokkuvõtete genereerimine	-	Uute päringutega katsetamine (hallutsineerimise vähendamiseks)	10-30 tundi (olenevalt päringute arvust)
Taotlustele hindajate leidmine	-	Taotlustele sobivate hindajate leidmise analüüs	25 tundi
Taotlustele hindajate leidmine	-	Andmete hõivamine OpenAlexist (miljon dokumenti)	60 tundi
Taotlustele hindajate leidmine	-	OpenAlexi andmete töötlemine ja vektoriseerimine sarnasusotsingus kasutamiseks (miljon dokumenti)	60 tundi
Taotlustele hindajate leidmine	-	Huvide konfliktiga hindaja välistamine analüüs	25 tundi
Taotlustele valdkondade määramine	ETIS projektid	Tagger Group otsingute rakendamine	15 tundi ⁴⁴
Arengukavade vastavusanalüüsi abiline / ETIS otsingute täiendamine / Taotlustele hindajate leidmine	Arengukavad jt andmed	Sarnasusotsingud: rakenduse arendus	160 tundi
Üldine arendus		Sarnasusotsingud: autentimine	15 tundi
Üldine arendus	Arengukavad jt andmed	Sarnasusotsingud: testimine	15 tundi

⁴⁴ Maht vabavaralise Texta Tookiti kasutamisel.

ETIS otsingute täiendamine	-	Lemmatiseerimise rakendamine	15 tundi ³⁵
Populaarteaduslike kokkuvõtete genereerimine	-	LLM-mudelite kaudu kokkuvõtete genereerimine: rakenduse arendus	55 tundi
Üldine arendus	-	LLM-mudelite kaudu kokkuvõtete genereerimine: testimine	15 tundi
Üldine arendus	-	Hostimine	Olenevalt andmemahust ja hostimise kestusest.

11. Lisad

Lisa 1. LLM-i abil valdkondade määramine, esimene päring

Väljamõeldud valdkonnad on märgitud helepunase värviga, vähemalt korra õigesti määratud valdkonnad helerohelise värviga ning ChatGPT õigesti määratud, kuid OpenAlexis teise valdkonnaga kollase värviga.

doi	answer_1	answer_2	answer_3	OA primary	OA all fields
http://doi.org/10.1109/TNS.2008.924077	Spectroscopy	Radiological and Ultrasound Technology	Radiological and Ultrasound Technology	Materials Chemistry	['Materials Chemistry']
http://dx.doi.org/10.1515/fca-2017-0058	Numerical Analysis	Numerical Analysis	Computational Mathematics	Modeling and Simulation	['Modeling and Simulation', 'Numerical Analysis']
https://doi.org/10.1163/9789004545687_004	Religious studies	Religious studies	Religious studies	Sociology and Political Science	['Sociology and Political Science']
http://dx.doi.org/10.1021/acscami.7b11044	Condensed Matter Physics	Condensed Matter Physics	Condensed Matter Physics	Materials Chemistry	['Materials Chemistry', 'Catalysis']
http://doi.org/10.1109/ISSPIT.2008.4775694	Computer Vision and Pattern Recognition	Computer Vision and Pattern Recognition	Computer Vision and Pattern Recognition	Signal Processing	['Signal Processing']
https://doi.org/10.1021/jacs.3c09551	Organic Chemistry	Organic Chemistry	Organic Chemistry	Organic Chemistry	['Organic Chemistry']
http://doi.org/10.1016/j.jpccs.2016.05.008	Catalysis	Catalysis	Catalysis	Mechanical Engineering	['Mechanical Engineering', 'Materials Chemistry', 'Catalysis']

http://doi.org/10.1016/j.cej.2008.06.013	Materials Chemistry	Metals and Alloys	Metals and Alloys	Mechanical Engineering	['Mechanical Engineering', 'Materials Chemistry']
http://doi.org/10.1016/j.bandc.2013.09.006	Experimental and Cognitive Psychology	Experimental and Cognitive Psychology	Experimental and Cognitive Psychology	Cognitive Neuroscience	['Cognitive Neuroscience', 'Experimental and Cognitive Psychology']
https://doi.org/10.1484/J.ML.C.5.116540	Archeology	Archeology	Archeology		[]
http://doi.org/10.1080/15568318.2013.821003	Transportation	Sociology and Political Science	Sociology and Political Science	Transportation	['Transportation']
http://dx.doi.org/10.1016/j.chemosphere.2022.134182	Nanotechnology	Materials Chemistry	Environmental Chemistry	Materials Chemistry	['Materials Chemistry']
http://dx.doi.org/10.37500/IJESSR.2020.30216	Development	Educational Psychology	Educational Psychology	Education	['Education']
http://doi.org/10.1071/978-981-10-5122-7_5	Human Factors and Ergonomics	Psychiatry and Mental Health	Psychiatry and Mental Health	Cognitive Neuroscience	['Cognitive Neuroscience']
http://dx.doi.org/10.1016/j.tels.2015.01.001	Sociology and Political Science	Sociology and Political Science	Information Systems and Management	Statistical and Nonlinear Physics	['Statistical and Nonlinear Physics', 'Information Systems and Management']

http://doi.org/10.1016/j.techfore.2017.05.025	Sociology and Political Science	Economics and Econometrics	Economics and Econometrics	Information Systems	['Information Systems', 'Sociology and Political Science', 'Marketing']
http://dx.doi.org/10.1080/09540962.2015.1047264	Literature and Literary Theory	Literature and Literary Theory	Literature and Literary Theory	Sociology and Political Science	['Sociology and Political Science']
http://doi.org/10.1080/13876988.2015.1129737	Political Science and International Relations	Political Science and International Relations	Political Science and International Relations	Economics and Econometrics	['Economics and Econometrics', 'Political Science and International Relations']
https://doi.org/10.1057/978-1-137-55269-3_56	Political Science and International Relations	Political Science and International Relations	Political Science and International Relations	Political Science and International Relations	['Political Science and International Relations', 'Public Administration', 'Economics and Econometrics']
http://doi.org/10.1177/1024258916683865	Sociology and Political Science	Management	Sociology and Political Science	Sociology and Political Science	['Sociology and Political Science', 'Strategy and Management', 'Management of Technology and Innovation']
https://doi.org/10.1016/B978-0-12-804137-6.00013-9	Management	Management	Management	Political Science and International Relations	['Political Science and International Relations', 'Management of Technology and Innovation', 'Economics and Econometrics']

Lisa 2. LLM-i abil valdkondade määramine, teine päring

Väljamõeldud valdkonnad on märgitud helepunase värviga, vähemalt korra õigesti määratud põhivaldkonnad helerohelise värviga ning kõigi valdkondade seast määratud õiged valdkonnad kollase värviga.

DOI	Vastus 1	Vastus 2	Vastus 3
http://doi.org/10.1109/TNS.2008.924077	Material Science	Materials Science	Materials Science
http://dx.doi.org/10.1515/fca-2017-0058	Numerical Analysis	Numerical Analysis	Numerical Analysis
https://doi.org/10.1163/9789004545687_004	Religious Studies	History of Religion	Political Philosophy
http://dx.doi.org/10.1021/acsami.7b11044	Materials Science	Materials Science	Materials Science
http://doi.org/10.1109/ISSPIT.2008.4775694	Computer Vision	Computer Science	Computer Vision
https://doi.org/10.1021/jacs.3c09551	Organic Chemistry	Organic Chemistry	Organic Chemistry
http://doi.org/10.1016/j.jpcs.2016.05.008	Materials Science and Catalysis	Materials Science	Material Science and Catalysis
http://doi.org/10.1016/j.cej.2008.06.013	Materials Science	Materials Science	Materials Science
http://doi.org/10.1016/j.bandc.2013.09.006	Neuroscience	Cognitive Neuroscience	Cognitive Neuroscience
http://doi.org/10.1080/15568318.2013.821003	Transport studies	Transport Studies	Transport Studies
http://dx.doi.org/10.1016/j.chemosphere.2022.134182	Materials Science	Materials Science	Materials Science
http://dx.doi.org/10.37500/IJESSR.2020.30216	Educational Psychology	Educational Psychology	Educational Psychology
http://doi.org/10.1007/978-981-10-5122-7_5	Occupational Health / Neuroscience	Neuroscience.	Neuroscience

http://dx.doi.org/10.1016/j.tele.2015.01.001	Information Science	Information Science	Library and Information Science
http://doi.org/10.1016/j.techfore.2017.05.025	Economics	Blockchain Technology in Economics.	Economics
http://dx.doi.org/10.1080/09540962.2015.1047264	Humanities	Philosophy	Social Sciences
http://doi.org/10.1080/13876988.2015.1129737	Political Science	Public Administration and Policy	Political Science
https://doi.org/10.1057/978-1-137-55269-3_56	Public Administration	Public Administration	Public Administration
http://doi.org/10.1177/1024258916683865	Social Sciences	Social Economy and Cooperative Studies	Economics
https://doi.org/10.1016/B978-0-12-804137-6.00013-9	Policy Studies	Public Policy and Governance	Public Policy and Governance

Lisa 3. LLM-i abil valdkondade määramine, kolmas päring

Väljamõeldud valdkonnad on märgitud helepunase värviga, vähemalt korra õigesti määratud põhivaldkonnad helerohelise värviga ning kõigi valdkondade seast määratud õiged valdkonnad kollase värviga.

DOI	Vastus 1	Vastus 2	Vastus 3
http://doi.org/10.1109/TNS.2008.924077	Condensed Matter Physics	Spectroscopy	Spectroscopy
http://dx.doi.org/10.1515/fca-2017-0058	Computational Mathematics	Numerical Analysis	Computational Mathematics
https://doi.org/10.1163/9789004545687_004	Religious studies	Religious studies	Religious studies
http://dx.doi.org/10.1021/acsami.7b11044	Condensed Matter Physics	Condensed Matter Physics	Condensed Matter Physics
http://doi.org/10.1109/ISSPIT.2008.4775694	Computer Vision and Pattern Recognition	Computer Vision and Pattern Recognition	Computer Vision and Pattern Recognition

https://doi.org/10.1021/jacs.3c09551	Organic Chemistry	Organic Chemistry	Organic Chemistry
http://doi.org/10.1016/j.ipcs.2016.05.008	Catalysis	Catalysis	Catalysis
http://doi.org/10.1016/j.cej.2008.06.013	Materials Chemistry	Materials Chemistry	Materials Chemistry
http://doi.org/10.1016/j.bandc.2013.09.006	Experimental and Cognitive Psychology	Experimental and Cognitive Psychology	Experimental and Cognitive Psychology
http://doi.org/10.1080/15568318.2013.821003	Transportation	Transportation	Transportation
http://dx.doi.org/10.1016/j.chemosphere.2022.134182	Applied Microbiology and Biotechnology	Applied Microbiology and Biotechnology	Applied Microbiology and Biotechnology
http://dx.doi.org/10.37500/IJESSR.2020.30216	Education	Developmental and Educational Psychology	Educational Psychology
http://doi.org/10.1007/978-981-10-5122-7_5	Occupational Therapy	Organizational Behavior and Human Resource Management	Occupational Therapy
http://dx.doi.org/10.1016/j.tele.2015.01.001	Information Systems	Communication, Research and Theory	Human-Computer Interaction
http://doi.org/10.1016/j.techfore.2017.05.025	Management of Technology and Innovation	Economics and Econometrics	Economics and Econometrics
http://dx.doi.org/10.1080/09540962.2015.1047264	Philosophy	Philosophy	Philosophy
http://doi.org/10.1080/13876988.2015.1129737	Public Administration	Public Administration	Public Administration
https://doi.org/10.1057/978-1-137-55269-3_56	Public Administration	Public Administration	Public Administration
http://doi.org/10.1177/1024258916683865	Sociology and Political Science	Management Science and Operations Research	Organizational Behavior and Human Resource Management

https://doi.org/10.1016/B978-0-12-804137-6.00013-9	Management, Monitoring, Policy and Law	Strategy and Management	Strategy and Management
---	--	----------------------------	----------------------------

Lisa 4. Tagger Group API juhend

Et Tagger Groupi kasutada tuleb end Texta Toolkitis registreerida. Registreerimiseks minna [Texta Toolkit](#) lehele ning luua seal kasutaja kasutades **Register** nuppu või APIs <https://rest.texta.ee/api/v2/rest-auth/registration/>. Loodud kasutajanime ning parooli läheb vaja APIs autentimiseks.

Autentimiseks tuleb genereerida token <https://rest.texta.ee/api/v2/rest-auth/login/> kaudu. Registreerimisel luuakse unikaalne token, mille abil saab APIs Tagger Groupidele päringuid teha (järgnevas näites "XXXX"). Vaata rohkem: <https://docs.texta.ee/authentication.html#api>

Valdkondade määramiseks kasutatakse Tagger Group API-t:

```
curl -X POST
"https://rest.texta.ee/api/v2/projects/355/tagger_groups/481/tag_text/" \
-H "accept: application/json" \
-H "Content-Type: application/json" \
-H "Authorization: Token XXXX" \
-d '{
    "text": "Typical tendencies in thinking, feeling, and behaving can be
summarised by personality and cognitive traits and measured with tests. These
behavioural traits associate with life outcomes, such as health and well-being.
However, causal directions of these associations are largely unknown, as
randomised controlled trials are resource-intensive and often ethically
impossible. We propose using a genomic causal inference method - Mendelian
randomisation. This method capitalises on natural randomisation of genetic
variants - genetic lottery - that cause differences in behavioural traits. To find
these variants, we propose collecting detailed behavioural data from 220,000
genotyped participants of the Estonian Biobank. The result will be an atlas of
plausibly causal associations between behaviour and life outcomes. The compiled
dataset will also be a valuable public resource for any researcher interested in
the intersection of behaviour, genetics, and health research for many years to
come."
    ,
    "lemmatize": false
}'
```

API annab vastuseks kõik võimalikud valdkonnad:

```
[
  {
    "tag": "Genetics",
    "probability": 0.6951252777203066,
    "tagger_id": 305635,
    "ner_match": false,
    "lexicon_id": null,
    "result": true
  },
  {
    "tag": "Clinical Psychology",
    "probability": 0.5888074399487326,
    "tagger_id": 305827,
    "ner_match": false,
    "lexicon_id": null,
    "result": true
  },
  {
    "tag": "Experimental and Cognitive Psychology",
    "probability": 0.5819098639046703,
    "tagger_id": 305833,
    "ner_match": false,
    "lexicon_id": null,
    "result": true
  },
  {
    "tag": "Applied Psychology",
    "probability": 0.5490718178465358,
    "tagger_id": 306175,
    "ner_match": false,
    "lexicon_id": null,
    "result": true
  },
  {
    "tag": "Public Health, Environmental and Occupational Health",
    "probability": 0.5200011253101696,
    "tagger_id": 305743,
    "ner_match": false,
    "lexicon_id": null,
    "result": true
  }
]
```

Ingliseelne (OpenAlexi andmetel põhinev) Tagger Group:

https://rest.texta.ee/api/v2/projects/355/tagger_groups/481/tag_text/

Eestikeelne (ETIS projektide andmetel põhinev) Tagger Group:

https://rest.texta.ee/api/v2/projects/355/tagger_groups/475/tag_text/

Eestikeelse Tagger Groupi puhul palume kasutada päringus käsku `"lemmatize": true`.

API kohta rohkem informatsiooni: https://docs.texta.ee/tagger_group.html#tag-text

Lisa 5. Kokkuvõtete koostamise eksperimendis kasutatud publikatsioonide viited

<http://doi.org/10.5128/ERYa19.08>

<http://doi.org/10.5128/ERYa13.02>

<http://doi.org/10.12697/jeful.2016.7.2.02>

<http://doi.org/10.5128/ERYa13.07>

<http://doi.org/10.12697/jeful.2016.7.2.05>

<http://dx.doi.org/10.5128/ERYa17.06>

<http://doi.org/10.22601/PET.2019.04.03>

<http://doi.org/10.5128/ERYa16.02>

<http://doi.org/10.5128/ERYa13.12>

<https://doi.org/10.5128/ERYa20.10>

<http://doi.org/10.2112/JCOASTRES-D-10-00049.1>

<http://doi.org/10.12697/jeful.2013.4.3.05>

<https://doi.org/10.5128/ERYa19.07>

<http://dx.doi.org/10.5128/ERYa14.11>

<http://dx.doi.org/10.5128/ERYa17.08>

<http://doi.org/10.22601/PET.2019.04.05>

<https://www.etis.ee/Portal/Publications/Display/cdea767c-02ba-4aa6-85ae-76555cd7fbf8>

<https://www.etis.ee/Portal/Publications/Display/d5d34334-4be5-45f2-8054-caf9906660d4>

<https://www.etis.ee/Portal/Publications/Display/fee0055e-de8a-43ac-82e4-1520e9d24b21>

<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8558456/pdf/fmicb-12-738125.pdf>

<https://www.sciencedirect.com/science/article/pii/S0304394009015511?via%3Dihub>

<https://pubmed.ncbi.nlm.nih.gov/33957149/>

<https://pubmed.ncbi.nlm.nih.gov/32348809/>

Lisa 6. LLM päringu publikatsiooni annotatsioon ning sellele loodud kokkuvõte

Publikatsioon: "Artiklis keskendutakse eesti keele polüseemilise ja sageda kasutusega tuumverbi laskma alustamis- (laskma + Vma, nt lasta vesi kraanis jooksmas) ja põhjustamiskonstruktsioonile (laskma + Vda, nt ei lasknud sind/sul ära sõita) ning kausatiivse funktsiooni kujunemise tagamaadele. Funktsiooni alltüüpidega käsitletakse permissiivset ja kuratiivset kausatiivsust. Kirjeldatakse laskma-verbi kasutusi kirjakeele tekstide põhjal alates XVI sajandist kuni tänapäevani ning selgitatakse tähenduseseid. Sarnase metoodikaga olen uurinud panema- ja ajama-verbi grammatiseerumist (Tomson 2018, 2020). Laskma-, ajama- ja panema-kausatiivkonstruktsioonid on grammatiseerunud eri tüüpi leksikaalsetest kasutustest. Põhjustamiskonstruktsioonina käsitletakse grammatilist konstruktsiooni, mis väljendab tervikuna olukorda, kus ühe sündmuse tagajärjel toimub teine sündmus (Comrie 1989; Langacker 1991; Talmy 2003). Analüütilise ehk perifrastilise kausatiivkonstruktsiooni keskmeks on liitpredikaat (nt paneb sädelema, ajab naerma, laseb laguneda) (Erelt, Metslang 2017: 231–236).

Laskma-verb esineb grammatilises konstruktsioonis da- või ma-infinitiiviga. Konstruktsiooni laskma + Vda (nt Ema laskis lapsel mängida) peetakse täpsemalt permissiiviks – põhjustamistarindi deontilis-modaalseks eriliigiks, mille kausatiivsus seisneb tulemusündmuse toimumise lubamises, selle mittetakistamises (Erelt, Metslang 2017: 231). Eesti keele laskma-verbil on nii lubav kui ka direktiivne (käsku, korraldust või palvet väljendav) tähendus (Erelt, Metslang 2017: 234). Kausatiivsust on ka teistes keeltes seostatud permissiivsusega (vt nt Diedrichsen 2015; Waldenfels 2015; Kuteva jt 2019: 256–257), nt inglise keele let-konstruktsioonil on võimaldamise tähendus: seda käsitletakse laiemalt kui kompleksset põhjustamissituatsiooni väljendavat konstruktsiooni, mis liigitatakse võimaldava kausatiivsuse (ingl enabling causation) alla. Seda tüüpi kausatiivsuse üks tunnuseid on see, et tulemusündmuse saavutamiseks ei rakendata tingimata füüsilisel kontaktil põhinevat jõudu (Talmy 2003: 504–509). Lisaks on osa laskma + Vda konstruktsiooni kasutusjuhtude puhul tegemist kuratiivse kausatiivsusega. Kui permissiivse kausatiivsuse puhul on mõjutaja (põhjustaja) protsessi või tegevuse mittetakistaja (passiivne lubaja) ja mõjutatav võib olla kolmes eri rollis: AGENT, KOGEJA või PATSIENT, siis kuratiivse kausatiivsuse puhul on nii mõjutajal kui ka mõjutataval aktiivne roll (Kasik 2001: 111). Kuratiivse kausatiivsuse fookuses on millegi saavutamine (Waldenfels 2015). Laskma + Vma konstruktsioon väljendab aga aspektitähendust, täpsemalt millegi alustamist.

Grammatiliste konstruktsioonide kujunemist vaadeldakse kui loomulikku grammatiseerumisprotsessi, milles leksikaalsest keeleainesest tekib grammatikaüksusi (Heine jt 1991; Hopper, Traugott 2003; Narrog, Heine 2011; Heine 2018; Kuteva jt 2019), aga ka kui kontaktimõjulist grammatiseerumist, sh sundgrammatiseerumist (Nau 1995; Heine, Kuteva 2005). Eesti oludes ilmneb see kõige ilmekamalt vana kirjakeele perioodil alates XVI sajandist: laskma-konstruktsiooni kasutati tõlkelises XVI–XVII sajandi kirjakeeles saksa keele eeskujul, näiteks Heinrich Stahli tekstides (VAKK). Kuna XVIII sajandi näidetest on näha, et

laskma-konstruktsioon sobitus hästi eesti keelde, siis võib arvata, et samal ajal on toimunud loomulik grammatiseerumine.

Artiklis avatakse seoseid laskma-verbi leksikaalsete kasutuste, sildkonstruktsioonide (leksikaalsete ja grammatiliste kasutuste vahepealsed, mitmeti tõlgendatavad juhud, mille kaudu saab selgitada grammatiseerumist) ning grammatiliste kasutuste vahel. Täendusrühmadesse jagatakse üle 3000 laskma-verbi kasutusjuhu, jälgides sajandite kaupa nende tähenduste kirjalikes tekstides avaldumise sagedust. Täendusseoste uurimise kaudu on eesmärk heita valgust grammatiliste konstruktsioonide kujunemisele. Grammatilistest konstruktsioonidest käsitletakse lähemalt da-infinitiiviga konstruktsioone käskimise (nt ma lasen sulle kohvi siia tuua) (direktiivsed kasutused) ja võimaldamise (nt oleme lasknud neil elada teiste arvel) (permissiivsed kasutused) tähenduses ning ma-infinitiivilist konstruktsiooni alustamistähenduses (nt lasi ringi käima lugemissoovituste nimestikud). Konstruktsiooni osaliste semantiliste rollide kaudu tuuakse välja ka sagedamini esinevad laskma-verbi sisaldava põhjustamiskonstruktsiooni tüübid.⁴⁵

Kokkuvõte: "Artikkel keskendub eesti keele sõna "laskma" mitmekesise kasutuse uurimisele. "Laskma" on "tuumverb" ehk sellel on palju tähendusi ning seda kasutatakse tihti. Laskma põhiline kasutus on alustamis- ja põhjustamiskonstruktsioonides.

Artiklis uuritakse, kuidas sõna "laskma" on erinevates konstruktsioonides kasutatud alates 16. sajandist kuni tänapäevani. Kasutatakse kahte peamist tüüpi konstruktsiooni:

1. Alustamiskonstruktsioon: See tähistab tegevuse alustamist, näiteks "lasta vesi kraanis jooksmas."
2. Põhjustamiskonstruktsioon: See väljendab tegevuse lubamist, takistamist või põhjustamist, näiteks "ema laskis lapsel mängida" või "ei lasknud mul ära sõita."

Neid kahte eristab teise tegusõna lõpp. Alustamiskonstruktsioonis on tegusõna lõpp -ma, põhjustamiskonstruktsioonis -da.

Põhjustamiskonstruktsioone liigitatakse edasi kolmeks:

- Permissiivne kausatiivsus: Lubamine või mitte takistamine, kus mõjutaja laseb millelgi juhtuda ilma otseselt sekkumata, samuti võib see tähendada käsku või palvet. Näiteks "ema laskis lapsel mängida," kus ema lubab lapsel mängida, kuid ei suuna otseselt tema tegevust. Tulemuse saavutamiseks ei rakendata tingimata füüsilisel kontaktil põhinevat jõudu. Sarnaseid konstruktsioone on ka teistes keeltes, nt inglise keeles *let*-konstruktsioon.
- Kuratiivne kausatiivsus: Aktiivsemat rolli eeldav lubamine, kus nii mõjutaja kui ka mõjutatav on aktiivsed. Näiteks "ma lasin arstil end ravida". Mõlemal osapoolel on aktiivne roll ja fookus on millegi saavutamisel.
- Direktiivne kausatiivsus: Käskimise tähendust omav funktsioon, näiteks "ma lasen sulle kohvi siia tuua" viitab sellele, et kellelgi antakse suunis, kuidas midagi teha.

Artikkel uurib sõna "laskma" grammatikaliseerumist. Grammatikaliseerumine on teooria, kuidas ühe tähendusega sõnast saab aja jooksul grammatika osa, millele tekivad algsest teistsugused

⁴⁵ Tomson, Kairit 2024. Laskma-verbi sisaldavad grammatilised konstruktsioonid ja nende tausttähendused eesti kirjakeeles. Keel ja Kirjandus 2024 (6).

tähendused ja funktsioonid. Grammatikaliseerumine võib tekkida iseenesest või tulla välismõjudest. Eestis kasutati alates 16. sajandist laskma-konstruksiooni saksa keele eeskujul, näiteks Heinrich Stahli tekstides. Et 18. sajandi näidetest on laskma-konstruksioon hästi eesti keelde kasutusse jäänud, siis võib arvata, et samal ajal on toimunud loomulik grammatiseerumine. Ehk siis on alguses grammatilised funktsioonid tulnud saksa keele mõjust, uued funktsioonid on kinnistunud ja loomulikult edasi arenenud.

Uuringus jagatakse rohkem kui 3000 "laskma"-verbi sõnavaralist, grammatilist ning vahepealset kasutusjuhtu tähenduse järgi gruppidesse, et analüüsida nende kasutamise sagedust ja tähenduse muutusi läbi sajandite. Uuringu eesmärk on näidata, kuidas grammatilised konstruksioonid on välja kujunenud."

Lisa 7. Projekti annotatsioonile ja valdkonnale sarnaseimad publikatsioonid

Sarnasus-skoor	DOI	Leitud sisukokkuvõte
1.7170476	https://doi.org/10.20506/rst.27.2.1820	Climate change is predicted to have important effects on parasitism and disease in freshwater and marine ecosystems, with consequences for human health and socio-economics. The distribution of parasites and pathogens will be directly affected by global warming, but also indirectly, through effects on host range and abundance. To date, numerous disease outbreaks, especially in marine organisms, have been associated with climatic events such as the El Niño-Southern Oscillation. In general, transmission rates of parasites and pathogens are expected to increase with increasing temperature. Evidence suggests that the virulence of some pathogens and parasites may also increase with global warming. The effects of climate change on parasites and pathogens will be superimposed onto the effects of other anthropogenic stressors in ecosystems, such as contaminants, habitat loss and species introductions. This combination of stressors may work cumulatively or synergistically to exacerbate negative effects on host organisms and populations. Climatic effects on parasites and diseases of key species may cascade through food webs, with consequences for entire ecosystems.
1.7161313	https://doi.org/10.1177/1178633617732296	The connection between our environment and parasitic diseases may not always be straightforward, but it exists nonetheless. This article highlights how climate as a component of our environment, or more specifically climate change, has the capability to drive parasitic disease incidence and prevalence worldwide. There are both direct and indirect implications of climate change on the scope and distribution of parasitic organisms and their associated vectors and host species. We aim to encompass a large body of literature to demonstrate how a changing climate will perpetuate, or perhaps exacerbate, public health issues and economic stagnation due to parasitic diseases. The diseases examined include those caused by ingested protozoa and soil helminths, malaria, lymphatic filariasis, Chagas disease, human African trypanosomiasis, leishmaniasis, babesiosis, schistosomiasis, and echinococcus, as well as parasites affecting livestock. It is our goal to impress on the scientific community the magnitude a changing climate can have on public health in relation to parasitic disease burden. Once impending climate changes are now upon us, and as we see these events unfold, it is critical to create management plans that will protect the health and quality of life of the people living in the communities that will be significantly affected.

1.6941407	https://doi.org/10.1016/s0020-7519(00)00141-7	<p>Ecological disturbances exert an influence on the emergence and proliferation of malaria and zoonotic parasitic diseases, including, Leishmaniasis, cryptosporidiosis, giardiasis, trypanosomiasis, schistosomiasis, filariasis, onchocerciasis, and loiasis. Each environmental change, whether occurring as a natural phenomenon or through human intervention, changes the ecological balance and context within which disease hosts or vectors and parasites breed, develop, and transmit disease. Each species occupies a particular ecological niche and vector species sub-populations are distinct behaviourally and genetically as they adapt to man-made environments. Most zoonotic parasites display three distinct life cycles: sylvatic, zoonotic, and anthroponotic. In adapting to changed environmental conditions, including reduced non-human population and increased human population, some vectors display conversion from a primarily zoophyllic to primarily anthrophyllic orientation. Deforestation and ensuing changes in landuse, human settlement, commercial development, road construction, water control systems (dams, canals, irrigation systems, reservoirs), and climate, singly, and in combination have been accompanied by global increases in morbidity and mortality from emergent parasitic disease. The replacement of forests with crop farming, ranching, and raising small animals can create supportive habitats for parasites and their host vectors. When the land use of deforested areas changes, the pattern of human settlement is altered and habitat fragmentation may provide opportunities for exchange and transmission of parasites to the heretofore uninfected humans. Construction of water control projects can lead to shifts in such vector populations as snails and mosquitoes and their parasites. Construction of roads in previously inaccessible forested areas can lead to erosion, and stagnant ponds by blocking the flow of streams when the water rises during the rainy season. The combined effects of environmentally detrimental changes in local land use and alterations in global climate disrupt the natural ecosystem and can increase the risk of transmission of parasitic diseases to the human population.</p>
1.6795465	https://doi.org/10.1016/j.vetpar.2015.04.022	<p>Over the last few decades, the world has witnessed radical changes in climate, landscape, and ecosystems. These events, together with other factors such as increasing illegal wildlife trade and changing human behaviour towards wildlife, are resulting into thinning boundaries between wild canids and felids and their domestic counterparts. As a consequence, the epidemiology of diseases caused by a number of infectious agents is undergoing profound readjustments, as pathogens adapt to new hosts and environments. Therefore, there is a risk for diseases of wildlife to spread to domestic carnivores and vice versa, and for zoonotic agents to emerge or re-emerge in human populations. Hence, the identification of the hazards arising from the co-habitation of these species is critical in order to plan and develop adequate control strategies against these pathogens. In the first of this two-part article, we review the role that wild canids and felids may play in the transmission of protozoa and arthropod-borne agents to dogs and cats in Europe, and provide an account of how current and future progress in our understanding of the ecology and epidemiology of parasites, as well as of host-parasite interactions, can assist efforts aimed at controlling parasite transmission.</p>

1.678643	https://doi.org/10.1016/s0020-7519(00)00128-4	<p>The environmental route of transmission is important for many protozoan and helminth parasites, with water, soil and food being particularly significant. Both the potential for producing large numbers of transmissive stages and their environmental robustness, being able to survive in moist microclimates for prolonged periods of time, pose a persistent threat to public and veterinary health. The increased demands on natural resources increase the likelihood of encountering environments and produce contaminated with parasites. For waterborne diseases, the protozoa, Cryptosporidium, Giardia and Toxoplasma, are the most significant causes, yet, with the exception of Toxoplasma, the contribution of zoonotic transmission remains unclear due to the absence of 'standardised' methods. The microsporidia have been documented in one waterborne outbreak, but the role of animals as the cause of contamination was not elucidated. In foods, surface contamination is associated with the faecal-oral pathogens, and some data are available to indicate that animal wastes remain an important source of contamination (e.g. cattle faeces and apple cider outbreaks), however, further work should focus on examining the source of contamination on fruit and vegetables. Increasing recognition of the burden of human fascioliasis has occurred; it is now recognised as an emerging zoonosis by the WHO. Toxoplasma, Trichinella and Taenia spp. remain important meatborne parasites, however, others, including Pleistophora-like microsporidians may be acquired from raw or lightly cooked fish or crustaceans. With increased international travel, the public health importance of the foodborne trematodiasis must also be realised. Global sourcing of food, coupled with changing consumer vogues, including the consumption of raw vegetables and undercooking to retain the natural taste and preserve heat-labile nutrients, can increase the risk of foodborne transmission. A greater awareness of parasite contamination of our environment and its impact on health has precipitated the development of better detection methods. Robust, efficient detection, viability and typing methods are required to assess risks and to further epidemiological understanding.</p>
----------	---	--