

TEHISINTELLEKTI RAKENDAMINE ERIALAKEELE ARENDAMISEL RIIGIKAITSETERMINOLOOGIA NÄITEL

Sandra Eiche, Reet Hendrikson, Eleri Aedmaa,
Esta Prangel, Mari-Liis Tikerperi

Ülevaade. Suurte keelemodelite ja tehisintellekti kiire areng pakub võimalusi nende mitmekülgselt rakendamiseks, mh erialakeele arendamisel. Artikli eesmärk on analüüsida riigikaitsevaldkonna terminitöö näitel välistoelise genereerimise (RAG) sobivust ja tõhusust definitsioonide ja allikaviidete väljatoomisel alusdokumentidest. Esialgsed tulemused näitavad, et RAG-süsteem leiab alusandmetest täpselt definitsioone ja allikaviiteid, mis teeb sellest sobiliku abivahendi terminoloogi töö tõhustamiseks. Seega on alust eeldada, et RAG-süsteemi abil saab toetada teisigi terminitöö etappe, nt pakkuda mõistekirje koostamiseks välja vajalikku infot, aidata visandada mõistesüsteemi, viidata võimalikele probleemidele algtekstis jm. Süsteem ei suuda terminoloogi küll asendada, ent abivahendina on RAG terminiarenduses perspektiivikas.*

Võtmesõnad: välistoeline genereerimine (RAG), suured keelemudelid, erialakeel, terminoloogia, definitsioon, mõiste, termin, riigikaitse, sõjateadus, eesti keel, inglise keel

1. Sissejuhatus ja eesmärk

Terminoloogia kui teadusharu ja erialakeele arendamise kontekstis kõneldakse tehisintellektil (TI) põhinevate rakenduste kasutuselevõtuga seoses uue ajajärgu algusest (vt nt Massion 2024: 92). Suurte keelemodelite (SKM) ja üldiselt generatiivse TI kasutuselevõtt nõuab terminiarenduse praktilise poole kõrval ka alus- seisukohtade ülevaatamist (Massion 2024: 92). Öeldu on loogiline, sest terminitöö eeldab andmemassiivide läbitöötamist. On vaid aja küsimus, millal generatiivsed süsteemid saavad igapäevase terminitöö ja terminihalduskeskkondade arendamise lahutamatuks osaks (vrd Reineke 2023: 28).

* Artikli valmimist on toetanud Kaitseministeeriumi teadus- ja arendusprojekt "Tehisintellekti rakendamine riigikaitseterminoloogia valdkonnas" (R-020).

TI roll inime töö hõlmajana on kümme aasta jooksul hüppeliselt suurenenud pea kõigil elualadel, seda ka riigikaitstes. NATO liikmesriigina järgib Eesti mh ka terminiarenduses NATO üldsuundi. NATO uusimad lähtedokumendid rõhutavad vajadust kasutada sõjalise eelise saamiseks üha enam TI pakutavaid võimalusi (NATO AI-strateegia 2024). Põhjus on lihtne: sõjapidamine on olnud alati seotud püüdega saavutada vastase suhtes tehnoloogilist ülekaalu. On üldteada fakt, et märkimisväärne osa teaduslikest ja tehnoloogilistest avastustest-uuendustest on kasutusele võetud sõdade käigus (Wright 1941, Mills, Mills 2014, Gross, Sampat 2020, Strickland 2023). Seetõttu on tänapäeva sõjaline võimearendus lahutamatu seotud TI-ga (NATO AI-strateegia 2024). Kaitseministeeriumi valitsemisala teadus- ja innovatsioonipoliitika (2022: 4) nimetab ühe eesmärgina teadus- ja arendustegevuste tõhustamist uuenduste lahendustega, mis võimaldab mh kiirendada andmeanalüüsi. See eesmärk on laiendatav kõikidele valdkondadele, sh erialakeele arendamisele.

Siinne artikkel kajastab pilootprojekti “Tehisintellekti rakendamine riigikaitse terminoloogia arendamisel” vahetulemusi ja tekkinud mõtte- ja kitsaskohti. Projekti eesmärk on välja töötada terminitööd toetava, TI-l põhineva tööriista prototüüp, et vabastada inime tööjõudu loovamaks, sisuliseks tööks. TI-lahendustest võib tõhusa rakendamise korral tuge saada terminitöö paljudes etappides, sh definitsioonide kontekstide otsimisel alg- ja/või sihtkeeles, mõistetunnuste selekteerimisel ja definitsioonide algversioonide loomisel sihtkeeles, võimalike sünonüümide leidmisel ja mõistesüsteemide visandamisel, ent ka juba terminibaasis sisalduvate probleemsete kirjade tuvastamisel. Neist tööloikudest on üks ajamahukamaid mõistete ja terminite alginfo leidmine andmekogudest. Nii sai projekti sihiks rakendada TI-d, et hõlpsamini leida termineid, definitsioone jm mõisteinfot, samuti teha alusandmetest aja- ja asjakohaseid väljavõtteid ning talletada seejuures allikaviited. Nii on võimalik suurendada uuendatud mõistekirjade ja -pesade hulka, säilitades senisega samaväärsed kvaliteedistandardid.

Riigikaitse sobib pilootprojektiks mitmetel põhjustel. Kiiresti arenevate valdkondade üks suuremaid väljakutseid on tulla toime uute mõistete rohke pealetungiga, mille tingib tehnika ja erialase mõtte kiire areng (vt Hendrikson 2018: 18–19, 2023). Ukraina sõda on nimetatud arenguid veelgi kiirendanud. On ilmne, et terminiarendus ei saa uute mõistete tulva kuidagi ennetada: alles uue mõiste lisandumisel saab hakata mõtestama selle sisu ja vahekorda naabermõistetega ning otsima optimaalseid omakeelseid vasteid. Väikekeele puhul tähendab uue mõiste lisandumine eeskätt seda, et vaste on laenuandvas keeles juba olemas ja mõiste on enamasti defineeritud ja/või mingil moel lahti seletatud. Lisaks peegelduvad riigikaitse terminitöös mitmed aspektid, mis aktualiseeruvad ka teiste valdkondade ja erialakeete puhul. Üks neist on keerukad mõistesuhted ja lähimõistete eristamise vajadus (Hendrikson 2014), mistõttu on arendatava prototüübi jaoks eraldi väljakutse võime sünonüümikandidaate välja pakkuda. Lisateemadena ilmnevad süsteemiga seotud turvakaalutlused, mh andmetöötlust ja hoiustamist puudutav, ning autorsus, mis terminisõnastikes väljendub allikaviidete olemasolus, süsteem-suses ja täpsusastmes.

Artikkel keskendub erialasõnastiku loomise ühele mahukamale tööloigule ehk mõistekirje algandmete leidmisele. Eesmärk on riigikaitse terminitöö näitel välja selgitada, kuivõrd kindlaks jääb RAG-süsteem definitsioonide ja allikaviidete

väljatoomisel alusandmetele. Otsitakse vastust küsimusele, kui hästi suudab RAG-süsteem leida definitsioone ja allikaviiteid. Esmalt käsitletakse lühidalt keeletehnoloogia ja TI senist kasutamist tõlkimisel ja sõnastike loomisel. Seejärel kirjeldatakse loodava RAG-süsteemi ülesehitust ja toimimispõhimõtteid ning analüüsitakse riigikaitse näidetele varal, kuidas süsteem sobib definitsioonide ja allikaviidete esitamiseks. Artikli lõpus esitatakse tulemused ja järeldused.

2. Keeletehnoloogia ja TI tõlkija ja sõnastikukoostaja abivahendina

Põhiosa terminisõnastike koostamist hõlbustavaid tööriistu on olnud siiani mõeldud eelkõige tõlkijate töö automatiseerimiseks. Arendatud on erinevatel keeletehnoloogilistel meetoditel rajanevaid terminihaldussüsteeme ja -tuvastajaid, mida on integreeritud mh tõlkesüsteemidesse. Tänu masintõlke kiirele arengule on lihtsustunud ka terminite tõlkeprotsess. (Massion 2024: 90) Vabavaralisena on näiteks kättesaadav mitmekeelseks termini- ja nende variantide tuvastuseks mõeldud TermSuite (Cram, Daille 2016: 13). Korpuspäringusüsteemis Sketch Engine (Kilgarriff jt 2014) on võtmesõnade funktsioon, mis võimaldab tuvastada korpusest termineid. Teiste hulgas ka eesti keeles töötava funktsiooni rakendamise eeldus on kvaliteetne valdkonnakorpus (Koppel, Kallas 2022: 220).

SKM-ide laialdane kasutamine kõikvõimalikel elualadel on kaasa toonud nende rakendamise ka keeleteaduses, nt korpuslingvistikas (Uchida 2024: 1). De Schryver (2023) tegi kümne teemakohase artikli ja ettekande põhjal kokkuvõtte ChatGPT kasutamisest leksikograafias ja leidis, et SKM-id võivad märkimisväärselt toetada sõnaraamatute koostamist, nt genereerida leksikograafidele kvaliteetset mustandmaterjali. Ta lisas, et produktiivsuse ja kvaliteedi tagamiseks tuleks välja töötada spetsiaalsed keelemudelid, kohandada olemasolevaid või kasutada neid konkreetsete alategevuste jaoks. (De Schryver 2023) Rees ja Lew (2024: 11–12) leidsid, et lugemisülesande lahendamisel ei sõltu keeleõppijate tulemused sellest, kas neile abivahendina esitada sõnastikus esitatud või TI genereeritud definitsioonid.

SKM-idel on hulk kitsaskohti, mida võiks püüda lahendada ja/või tulemuste analüüsimisel arvesse võtta, nt kalduvus hallutsineerida (ehk väljamõeldud vastuseid anda), läbipaistmatus (raske seletada, kuidas mudel teatud tulemuseni jõuab), mittedeterministlikkus (sama sisendi puhul ei saa alati sama väljundit) ja inglise keelele kallutatus (de Schryver 2023: 377–380). Lisaks tuuakse välja, et mudelite väljundites võib täheldada murettekitavaid stereotüüpe ning et mudeleid majutavatel andmekeskustel on halb keskkonnamõju (McKean, Fitzgerald, 2024).

Terminiarendusega seoses on uuritud, milliseid terminitöö etappe on otstarbekas TI-l põhinevate süsteemide abil kiirendada. Reineke (2023: 27) leiab, et suurema tehnilisusastmega valdkondades võib neid kasutades teha küll algul rohkem vigu (nt sünonüümide tuvastamine, mõistesüsteemi visandamine), ent tulemused annavad häid vihjeid edasiseks tööks. San Martín (2024) rõhutab, et valdkondades, kus täpsus ja järjepidevus on kriitilise tähtsusega, on hoolimata TI genereeritud sisu kasvavast mõjust inimeste loodud definitsioonid jätkuvalt asjakohased, sest need eeldavad inimestevahelist konsensust. Massion (2024: 92–93) toonitab, et generatiivse TI ajastul on oluline ära kasutada selle pakutavaid võimalusi terminitöö

paremaks korraldamiseks. Ometi vajab siin eraldi esiletoomist San Martíni tõdemus (2024), et terminisõnastiku loomisel võib tehnoloogiliste tööriistade kasutamine kujuneda sama töömahukaks kui kirje loomine TI toeta.

3. Suured keelemudelid ja RAG-i etapid

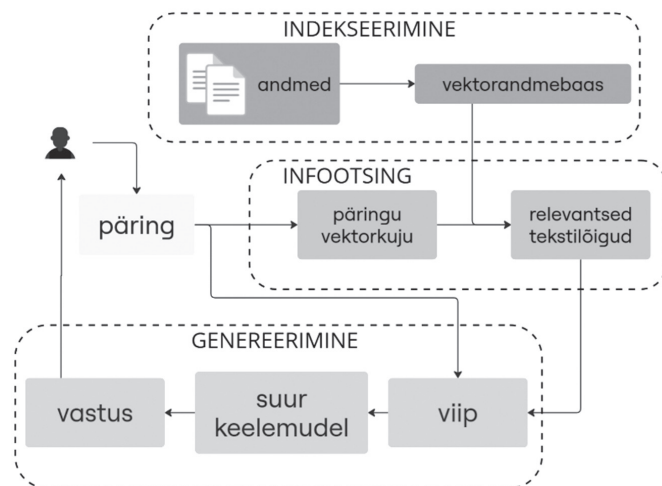
Kuigi SKM-id pakuvad juba praegusel kujul tuge erinevate töömahukate ülesannete lahendamisel, on asjaolusid, mis takistavad nende rakendamist (Zhang jt 2023). Näiteks ei saa eeldada, et laiatarbemudelid oskavad hallutsineerimata vastata mõnele kitsast valdkonda puudutavale küsimusele. Ebatäpse info edastamise vähendamiseks vajavad mudelid lisaandmeid või täiendavat juhendamist (Zhang jt 2023). Selleks tuleb esmalt tagada TI-le ligipääs erialastele andmetele ja teisendada need süsteemile loetavasse vormi.

Üks võimalus mudelile andmeid tutvustada on SKM-ide peenhäälestamine ehk olemasoleva mudeli treenimine mingit kindlat ülesannet lahendama kas juhendamise või lisaandmete abil (Fatehkia jt 2024). See lähenemine ei pruugi olla praktiline, sest vajalikke märkimisväärsed arvutusressursse (nt graafikaprotsessoreid) või piisavalt valdkonnaspetsiifilisi andmeid pole kuskilt võtta (Gao jt 2024). Lisaks võib olla tarvilik sisendandmete sage uuendamine, mis omakorda tähendab mudeli tihedat kohendamist (Fatehkia jt 2024).

Teine viis lisaandmete abil mudelite tulemusi parandada on kasutada välis-toelist genereerimist¹ (ingl *retrieval-augmented generation*, RAG) ehk meetodit, mis kasutab mudeli väljundi kvaliteedi parandamiseks andmeid välistest teadusallikatest (AKIT). Sellel meetodil põhinevad süsteemid ehk RAG-süsteemid leiavad esmalt päringu põhjal andmetest üles relevantssed lõigud, mille alusel koostatakse loogilised ja konteksti sobivad vastused (Lewis jt 2020: 9464). Nii vormistavad RAG-süsteemid lisaandmeid kasutades informatiivsemaid ja faktitäpsemaid vastuseid kui välise toeta SKM-id. Seetõttu leiab RAG aktiivset kasutust, mis omakorda on viinud üha paremini toimivate lahendusteni, mh on RAG-ist loodud mitmeid versioone, nt naiivne, edasijõudnud ja modulaarne. (Gao jt 2024) Edasises kesken-dutakse käsitletavas projektis rakendatud naiivsele RAG-ile, mida nimetatakse ka otsi-loe-raamistikuks (Ma jt 2023: 5303). Sel on kolm põhietappi: indekseerimine, infootsing ja genereerimine (vt joonis 1).

Indekseerimine hõlmab andmete eraldamist ehk parsimist, puhastamist, tükeldamist ja vektoriseerimist ning andmebaasi salvestamist (Gao jt 2024). Tükeldamisega jagatakse dokument väiksemateks osadeks, mis võimaldab hiljem terve dokumendi asemel vastata päringule lühemate ja relevantsemate tekstilõikudega. Heade tulemuste saamiseks on selles etapis tähtis arvestada, milliseid päringuid kasutaja tegema hakkab. Näiteks saab kasutada erinevaid tükeldamisstrateegiaid, millest lihtsaim ja levinum on sõnepõhine tükeldamine (ingl *token-based chunking*). See tähendab, et terviktekst jagatakse võrdseteks osadeks, millest igaüks sisaldab ettemääratud arvu märkide järjendeid ehk sõnesid (ingl *token*). Dokumente võib tükeldada ka lausete või pikemate lõikude kaupa. Pärast tükeldamist teisendatakse tekstilõigud eeltreenitud vektoriesituse sisaldava mudeliga (ingl *embedding model*) vektorkujule ning vektorid ja edasiseks kasutamiseks vajalikud metaandmed

32 ¹ Mõistet on Bogdanov jt (2024: 9) tähistanud terminiga *päringgenereerimine* ehk meetod, mille puhul keelemudel pärib andmebaasist vm välistest allikast kasutaja viiba põhjal täiendavat konteksti, et parandada vastuse kvaliteeti. Eestikeelses kirjanduses (vt nt Kivisalu 2024) esineb ka vaste *allikapõhine genereerimine*. Selles artiklis kasutatakse suupärasemat, ingliskeelsest terminist tuletatud lühendit RAG, mis on valdkonnaekspertide seas neist variantidest harjumuspäraseim.



Joonis 1. Naiivse RAG-süsteemi etapid (autorite koostatud
Lewise jt 2020, Gao jt 2024 põhjal)

salvestatakse baasi. Vektoriseerimine ja vektorandmebaaside kasutamine on RAG-süsteemi infootsingus kõige levinum lahendus. Alternatiivina kasutatakse klassikalisemaid märksõnapõhiseid loomuliku keele töötlemise vahendeid (Cuconasu jt 2024: 720).

Infootsingu tarbeks peab päringu vektoriseerima sama mudeliga, mida kasutati dokumentide vektoriseerimiseks. Päringu ja dokumentide vektorkujude võrdlemisel leitakse tekstilõigud, mis on päringuga kõige sarnasemad. Selleks rakendatakse vektoritevaheliste sarnasuste mõõtmiseks mõeldud sarnasusmõõdikut, nt koosinussarnasust. Otsingufaasis on kriitilise tähtsusega jõuda kõrge täpsuse ja saagiseni, et vältida valede või asjakohatute tekstilõikude valimist ning olulise teabe väljajätmist. (Gao jt 2024) Selleks, et tulemustesse jõuaks kõige relevantsem info, võib infootsingut täiustada päringu sõnastuse muutmise või info lisamisega.

Infootsinguga eraldatud tekstilõigud (edaspidi väljavõte) lisatakse SKM-ile saadetavale viibale (ingl *prompt*) ehk keelemudelilt tavakeeles küsitud küsimusele või päringule. See võib mh sisaldada probleemikirjeldust ja juhust, millisel kujul vastust oodatakse. Viiba põhjal koostabki SKM lõpliku vastuse, mis ideaalis tuleneb algandmetest, aga ei toetu neile liigselt, uut sisulist teavet andmata. Selles etapis võivad ilmned mudeli puudused, nt väljundi kahjustav või kallutatud sisu, mis vähendavad süsteemi usaldusväärsust. (Gao jt 2024)

4. RAG-süsteemi arhitektuur

Terminiarendust toetava prototüübi ülesehitamisel seab nõudeid ja piiranguid eeskätt see, kellele ja millisteks ülesanneteks süsteemi arendatakse. Rolli mängib mh valdkonna ja sisendinfo eripära. Praegusel juhul peab loodav prototüüp toetama erialasõnastiku koostaja tööd ning aitama andmeotsingul ja asjakohase info selekteerimisel.

Selleks on vaja esmalt leida mõistekirje jaoks infot algkeeles, st inglise keeles. Sisendinfost on vaja ekstraheerida võimalikud definitsioonivariandid, samuti leida selgitavaid kontekste, säilitades seejuures allikaviited. Terminisõnastiku loomise üks lähtekohti on, et kogu algkeelne info peab olema esitatud täpselt originaalkujul.

Teine põhimõte on, et kõik väljavõtted peavad olema varustatud allikaviitega. Riigikaitstes ja kitsamalt sõjanduses teeb olukorra keerukamaks see, et erinevad sõjanduskultuuriruumid (NATO, USA ja Suurbritannia) käsitlevad samu mõisteid veidi erinevalt. Kui üldmõistetes on erinevusi pigem vähe, siis relvaliigiti ja valdkonniti on need kohati märkimisväärsed. Siit tuleneb üks lisapõhjus, miks on allikaviidete olemasolu tähtis: see kannab infot ekspertide jaoks. Lisaks võivad erineda kasutatavad terminid, mistõttu on vajalik, et süsteem oleks võimeline välja pakkuma sünonüümikandidaate. Ekspertid otsustavad, millist definitsiooni kasutatakse, kas ja mida on vaja lisada märkuste väljale jne.

Kasutaja jaoks koosneb tööriist kolmest vaatest. Esimene on mõeldud huvi-pakkuvate dokumentide andmebaasi laadimiseks. Teises vaates saab kõigist aja jooksul üles laaditud dokumentidest teha semantilist otsingut ning kolmas vaade on sõnaraamatutest ja terminibaasidest ühispäringu tegemiseks.

4.1. Andmed

RAG-süsteemi arendamiseks kasutatud materjalides on erinevaid riigikaitselisi, peaaegselt sõjandusega seotud alusdokumente: riiklikud ja rahvusvahelised eeskirjad, seadused, raportid, manuaalid jms. Neid iseloomustab temaatiline variatiivsus: alates sõjanduse ja sõjapidamise põhimõtetest (sõjalist) riigikaitset käsitlevate dokumentideni. Materjalide valikul on arvestatud nende kasutust, valdkondlikku tähtsust ja asjakohasust, samuti dokumentide hierarhiat. Sisendandmetesse lisati ka ametliku kehtivuse kaotanud doktriine, kuna sealt on vahel võimalik leida mõiste kohta vajalikku diakroonilist infot.

Riigikaitse valdkonna eripärast lähtuvalt jääb osa dokumente kõrvale piiratud ligipääsu ja salastatuse tõttu. Praeguses arendusetapis on juurdepääsupiiranguga dokumendid sisendandmete seast välja jäetud. Edaspidi võiks olla üks arenduskohti seesuguste andmete lisamine, mis eeldab mh rangemaid reegleid süsteemi majutavale serverile, lokaalse mudeli kasutamist ja rollipõhise autentimislahenduse rakendamist. Teine, ehkki teistlaadi piirav tegur on asjaolu, et suurt osa ingliskeelsetest doktriinidest ja manuaalidest pole enam tarvidust eesti keelde tõlkida. Seega pole eestikeelseid sisendtekste nii palju, et neilt saaks väljundina oodata süsteemset mõisteinfo leidmist. Teisalt luuakse ohtralt eestikeelseid materjale, ent neil pole jällegi ingliskeelseid paralleeltekste. Leidub küll aastakümne(te) eest tõlgitud eestikeelseid materjale, millel on olemas ingliskeelsed originaalid, aga neis tõlkeis sisalduvatel terminitel on tänapäevaks suuresti vaid ajalooline väärtus.

4.2. Andmete üleslaadimine ja salvestamine

Dokumentides semantilise otsingu tegemiseks saab kasutajaliidese kaudu üles laadida uusi dokumente ja lisada neile metaandmeid, nt pealkiri, avaldamisaasta, märksõnad jm. Enne, kui kasutaja saab valitud dokumendist otsida, järgneb taustatööna info indekseerimine (vt ptk 3).

4.3. Dokumendipäringute ülesehitus

Dokumendipäringute fookusesse on võetud termini definitsioonide ja mõistepesadesse sobivate terminite leidmine. Praeguses etapis on kasutaja sisend termin ja lahendatavad ülesanded on defineeritud SKM-i viibas.

4.3.1. Info indekseerimine

Dokumendid jõuavad andmebaasidesse mitmes osas. Esmalt laetakse relatsioonilisse andmebaasi kasutaja sisestatud metaandmed, mida kasutatakse filtrite haldamiseks. Samal ajal parsitakse ja salvestatakse dokumendi sisu lehekülje kaupa koos metaandmetega faili, mille sisu samal ajal tükeldatakse ja vektoriseeritakse vektorandmebaasi tarbeks. Selles etapis toimub ka iga lõigu lehekülje tuvastamine.

Termin töö avastusliku olemuse tõttu sisaldab otsing ainult ühte terminit, mitte konkreetset küsimust. Seetõttu on oluline, et väljavõtted oleksid lühidamad ja sisult terviklikud, ent samas piisavalt pikad, et mahutaksid kogu olulise info. Siinse prototüübi arendamisel katsetati esmalt lausepõhist tükeldamismeetodit, kuid ilmnis, et sisenddokumendid sisaldasid sageli sealseid, mille lausestamise kvaliteet oli ebaühtlane. Andmete töötlemine sõnepõhise tükeldamisstrateegiaga andis rahuldavamaid tulemusi.

Andmed vektoriseeritakse mudeliga Multilingual-e5-Large, mis valiti mitmekeelsuse ja võimaluse tõttu teha lühemaid päringuid graafikaprotsessorita. Kuigi E5 mudelite sarjas (Wang jt 2024) on ka kiiremaid väiksema dimensiooniga mudeleid, valiti kõrgema täpsuse huvides 1024-dimensiooniline mudel. Suurem mudel sobib ka seetõttu, et uusi dokumente lisatakse baasi korraga vähe. Mitmekeelsus on oluline, et lubada eri keeltes päringuid ja sisenddokumente. Näiteks sisaldab kasutatud andmestik peale inglise keele ka vähesel määral prantsuse ja hispaania keelt.

4.3.2. Info pärimine

Kui dokumendid on vektorandmebaasis, saab alustada dokumendiotsingu vaates päringute tegemist. Esmalt on võimalik filtrimenüüst pealkirju või märksõnu selekteerides valida dokumendid, mille hulgast otsingut teostada. Kasutaja sisestab vestlusaknasse päringu, millega käivitub RAG-süsteem. Sisend vektoriseeritakse vektormudeliga, et arvutada koosinussarnasus baasis olevate dokumentidega ja seeläbi leida valitud andmetest kõige sarnasemad lõigud. Kuna erinevaid vektormudeleid lisandub pidevalt ja on võimalik, et kasutajate ootused tulevikus

muutuvad, on süsteem disainitud nii, et mudeli saab välja vahetada. Sel juhul tuleb arvestada, et uue mudeliga peab kõik andmed nullist vektoriseerima ja ka maksimaalne sisendipikkus või riistvaranõuded võivad olla praeguse mudeliga võrreldes erinevad.

4.3.3. Vastuse genereerimine

SKM-i vastus ilmub dokumendiotsingu vestlusaknas. Projekti skoobist, ressursidest ja terminoloogide vajadustest lähtuvalt kasutatakse siin OpenAI GPT-mudeleid. Samas on süsteem üles ehitatud nii, et lihtsasti saaks lisada toe ka teiste mudelite kasutamiseks.

Iga sisend ja selle põhjal infootsingust saadud väljavõte lisatakse viibale, mis kirjeldab seda, millistele küsimustele keelemudel vastama peab. Need võivad, aga ei pea põhinema infootsingul. Kuigi selles artiklis käsitletakse vaid definitsioonide sõnasõnalist andmetest eraldamist ja viitamist, on SKM-i abiga võimalik lahendada ka mitmeid teisi ülesandeid, nt seotud terminite esitamist. Lisaks vormistab SKM vastuse kasutajale mugavalt loetavale kujule. Praegu on süsteem sellises arendusetapis, mis ei võimalda süsteemilt lisaküsimusi küsida. Mudeli vastused ei ole deterministlikud, küll aga kattuvad ühel ja samal välisel andmestikul tehtud infootsingute vasted.

4.4. API-päringud

Peale RAG-süsteemi sisaldab prototüüp ka sõnaraamatutest ja terminibaasidest ühispäringu tegemise funktsiooni. Korraga saab teha märksõnaotsingu IATE terminibaasist ning sõnaraamatutest Collins English Dictionary, Collins Cobuild Advanced British, Collins Cobuild Advanced American ja Merriam-Webster. Neid allikaid kasutavad igapäevatöös Militermi koostajad ja nende andmed olid kättesaadavad avalike API-de ehk rakendusliidest kaudu. Kuigi tänu API-dele on võimalik infootsingule kuluvat aega kokku hoida, kaasneb nende kasutamisega ka proovikive. Näiteks sõltub päringute vastamisaeg suuresti nende teenusepakujatest ja kui otsingutulemusi on palju, võib tulemuste kuvamine võtta minuteid. Peale selle on allikate hulk piiratud, kuna kõik terminoloogid huvitavad terminibaasid ja sõnaraamatud ei paku avalikke rakendusliideseid või ei tee seda sobivatel tingimustel.

5. Hindamine

Arendatav RAG-süsteemi prototüüp koosneb mitmest komponendist, mille kvaliteet mõjutab omakorda süsteemi toimimist. Tööriista loomisel on olulised nii tehnilised kui ka sisulised aspektid, mille hindamiseks sobivad erinevad meetodid. Tavaliselt hinnatakse RAG-süsteemide puhul nii infootsingu- kui ka genereerimiskomponente. Infootsinguga seoses hinnatakse, kui täpselt ja asjakohaselt tuvastab süsteem dokumentidest infot, ning genereerimise hindamine annab vastuse küsimusele, kui täpne ja asjakohane on väljund kasutaja vaatest. Kuna projekti käigus valmiva

prototüübi põhjalikum hindamine on planeeritud 2024. aasta lõppu, siis pole praegu veel võimalik kõiki tulemusi ammendavalt kirjeldada, eriti neid, mis eeldavad valdkonnaekspertide põhjalikku sisulist tagasisidet. Küll aga on võimalik anda ülevaade sellest, kui kindlaks jääb süsteem andmetele ning mil määral esineb mudelil võimalike definitsioonide täpsel tuvastamisel ja viitamisel süsteemi usaldusväärsust kõigutavat hallutsineerimist. Selleks kõrvutati infootsingu ja genereerija väljundeid. Definitsioonide täpne eristamine on terminitöös vajalik, kuna terminibaasis tuleb algeelne teave esitada täpselt originaalkujul.

5.1. Valimi koostamine

Süsteemi hindamiseks kasutati riigikaitseterminite andmebaasi Militerm mõistekirjeid ja termineid. Militerm on pidevalt ajakohastatav terminibaas, mis koondab riigikaitsetermineid. Valim koosneb kirjetest, mille viimane sisuline muutmine toimus ajavahemikus 10.9.2019–9.9.2024 (k.a). Kirjetest eraldati ingliskeelsed terminid koos ingliskeelsete definitsioonidega. Välja jäeti kõik mitteavalikud terminid ja sellised avalikud terminid, mille väärtusolek oli “väldi”. Nii jäi alles 3168 terminit, millest osa esines Militermis rohkem kui ühes mõistekirjes. Duplikaatide eemaldamise järel jäi alles 2889 terminit, millele rakendati RAG-süsteemi.

Mudelile saadetud viiba² koostamise põhimõte oli, et väljundis esitatud definitsioon ei tohi olla mudeli interpretatsioon ega süntees, vaid peab edastama lähteandmed sõna-sõnalt. Siinkohal ei antud mudelile ette definitsiooni tunnuseid. Teiseks juhendati mudelit vastama kindlas formaadis, et tulemused oleksid hiljem automaatsel hindamisel paremini töödeldavad. Eesmärk oli jõuda võimalikult konkreetse, aga samas piisavat teavet sisaldava viibani. Selleks tuli läbida mitu testimisvooru, kasutades eri viipasid nii OpenAI odavaima mudeliga GPT-4o mini kui ka keerulisteks ülesanneteks mõeldud GPT-4o-ga. Tulemuste võrdlemise järel valiti hindamiseks GPT-4o. Süsteemiviibas, mida kasutatakse selleks, et mõjutada mudeli vastuste stiili ja vormistust, anti mudelile terminibaasi koostava terminoloogi roll. Mudelit suunati etteantud tekstidest definitsioone välja tooma, järgides vastuse andmisel kindlaid vormistusnõudeid, selle kohta lisati näited. Viiba alguses ja lõpus rõhutati, et definitsioonid peavad originaalteksti täielikult järgima. Sisestatud termin ja semantilise otsinguga leitud väljavõte edastati mudelile kasutajaviibas, et mudel saaks neile süsteemiviibas kirjeldatud moel vastata.

SKM-i vastuseid võrreldi automaatselt talle ette antud väljavõtetega. Automaatkontroll kasutas mudeli vastusest genereeritud definitsioonikandidaadi (kandidaat, sest mudeli vastus ei pruugi vastata defineerimisnõuetele ega olla sisuliselt sobilik) ning allikaviidete eristamiseks regulaaravaldisi, mille koostamisel eeldati, et keelemudeli vastus järgis viibas kirjeldatud vormistusjuhendit. Viite kaudu viidi väljavõttes kokku keelemudeli vastustega: kui väljavõte sisaldas vastust, loeti vastus korrektseks. Automaatselt loeti õigeks ka juhud, mil mudelile edastatud väljavõttes esines topelttühikuid, mida keelemudeli vastuses polnud.

² See viip ei ole sama sisuga, mis arendatava süsteemi prototüübi viip.

Tabel 1. Tulemused esialgsel ja lõppvalimil (* märgib käsitsi kontrolli järel saadud tulemusi)

Tunnus	Esialgne valim			Lõppvalim		
	kokku	leitud/ korrektsete arv	leitud/ korrektsete osakaal	kokku	leitud/ korrektsete arv	leitud/ korrektsete osakaal
Terminid	2889	2083	72,1%	350	–	–
Definiitsioon + viide	7717	7582	98,3%	1320	1304 1306*	98,8% 98,9%*
Kattuvus väljavõttega	7582	6527	86,1%	1304	1169 1221*	89,6% 92,5%*

Tabelis 1 on ülevaade tulemusest. 2889 terminist leiti definiitsioonikandidaat 2083-le. Teistele leidis süsteem vähemalt ühe, üldjuhul mitu definiitsiooni. 7717 definiitsioonikandidaadi-viite paarist olid korrektse viitega 98,3%. Neist omakorda 6527 (86,1%) kattus väljavõttega (võrdlust ei tehtud, kui allikaviide oli vale). Tulemuste lähemal uurimisel selgus, et automaatvõrdlus oli võimaliku definiitsiooni või viite lugenud valeks ka juhtudel, mis kasutaja vaatepunktist tundusid õiged. Sageli olid valenegatiivsed näited seotud PDF-ist parsimise eripäradega. Tulemuste valideerimiseks ja detailsemaks analüüsiks vähendati lõppvalimit 350-le³ juhuslikult valitud terminile.

5.2. Tulemused ja järeldused

Lõppvalimi 350 terminile leidis süsteem kokku 1320 definiitsioonikandidaadi-viite paari. Igal kandidaadil võis olla üks või rohkem allikaviidet. Mitu allikaviidet pidid saama sellised definiitsioonikandidaadid, mis esinesid täpselt samal kujul mitmes väljavõttes. Kokku andis süsteem 1320 viidet. Automaatkontrolliga tuvastati korrektne viide 1304 definiitsioonikandidaadil ja kuna neil juhtudel sai vastuse automaatselt väljavõttega siduda, on see hulk kattuvuse arvutamise alus. Definiitsioonikandidaat kattus väljavõttega 1169 korral (89,6%).

Väljavõtte ja definiitsioonikandidaadi erinevused (151) vaadati käsitsi üle ning selgus, et automaatkontroll oli erinevaks lugenud ka tekstikohti, kus erinevusel polnud sisulist tähtsust. Seesuguseid erinevusi võib liigitada järgmiselt.

1. Liigsed tühikud: sõnade vahele või sidekriipsu sisaldava liitsõna ja sidekriipsu vahele olid jäänud ebavajalikud tühikud, mille keelemudel oma vastuses eemaldas.
2. Kirjavahemärkide stiil: väljavõttes ja vastuses esinesid erinevad kirjavahe-märgistamise stiilid, nt ' ja `.
3. Kokkukäiva info ühendamise: mudel käsitas tervikuna väljavõtetes eraldatud, aga originaaldokumendis kõrvuti asetsenud lõike.
4. Poolikuks jäänud info: mudel markeeris väljavõttes poolikuks jäänud lause (nt kolme punktiga).
5. Lisainfo väljajätt: mudel jättis välja väljavõttes esinenud info, mis ei kuulunud definiitsiooni juurde, nt parsimise käigus sinna sattunud päise ja jaluse tekst.

³ Valimi vähendamine 350-ni tagab, et käsitsi töö maht on hallatav, aga valim on siiski piisavalt suur (valimi veapiir on 95% usaldusnivool 4,78%).

Mudel tegi vastuses ka selliseid parandusi, mis lähevad vastuollu viibas kirjeldatud definitsiooni sõnasõnalise väljatoomise nõudega, kuid samas viitavad sisendandmetes esinenud vigadele. Näites (1) jättis mudel originaaliga võrreldes välja sõna *red*, mida tõlgendas üleliigsena. Kasutaja vaatepunktist oleks tarvilik, et süsteem markeeriks vastuses ka hüpoteetilise weakoha, aga ise parandust ei teeks.

(1) Päring: ***light damage***

Väljavõte: [---] Damage which does not prevent the immediate use of equipment or installations for which it was intended. Some repair by the user may be required **red** to make full use of the equipment or installations. [---]

Definitsioonikandidaat: Damage which does not prevent the immediate use of equipment or installations for which it was intended. Some repair by the user may be required to make full use of the equipment or installations.

Alusandmetes on samad terminid sageli defineeritud täpselt ühtmoodi või väikse erinevusega, nt trükkivead, komakasutus ja regionaalsed erinevused. Viibas suunistati mudelit identse definitsioonikandidaadi puhul esitama seda vaid ühel korral ja järjestama selle all viited kõigile dokumentidele, kus see sellisel kujul esines. Vastuseid analüüsid selgus, et see tingimus oli mudeli jaoks keeruline. Mudel kippus sellistel juhtudel sisu ja sõnastuse põhjal üldistusi tegema ning edastas eelnevalt nimetatud väikeste vormiliste erinevustega definitsioonikandidaadid ühe ja samana (2).

(2) Päring: ***air supremacy***

Väljavõte: [---] Air supremacy is defined as that degree of air superiority wherein the opposing air **force is** incapable of effective interference. [---]

Definitsioonikandidaat: Air Supremacy is defined as that degree of air superiority wherein the opposing air **forces are** incapable of effective interference.

Näites (3), kus otsiti definitsiooni lühendile *EEI*, oli mudeli pakutud definitsioonikandidaadi ja väljavõtte vahel suuremad erinevused. Sellele vaatamata väitis mudel, et tegemist on ühe ja sama definitsiooniga. Näide (4) osutab, et erinevus võis seisneda isegi terves kõrvallauses.

(3) Päring: ***EEI***

Väljavõte: [---] The **most critical information requirements** regarding the **adversary** and the environment needed by the commander by a particular time to relate with other available information and intelligence in order to assist in reaching a logical decision. [---]

Definitsioonikandidaat: The critical items of information regarding the enemy and the environment needed by the commander by a particular time to relate with other available information and intelligence.

(4) Päring: ***line of departure***

Väljavõte: [---] In amphibious warfare, a suitably marked offshore coordinating line to assist assault craft to land on designated beaches at scheduled times. [---]

Definitsioonikandidaat: In amphibious operations, a suitably marked offshore coordinating line, **which is located at the seaward end of a boat lane**, to assist in the landing of landing craft and amphibious vehicles on designated beaches at the scheduled times.

Mõnel harval juhul oli väljavõtte ja definitsioonikandidaadi vahel vähe sarnasusi. Selle olukorra tekitas SKM-i võime mõista, et erinevad väljavõtted olid pärit sama dokumendi eri versioonidest. Kuna väljavõtetes oli osalisi kattuvusi, siis pakkus mudel, et sama definitsioonikandidaat esineb ka teises dokumendis, mille väljavõttest see tegelikult puudus. Esines ka juhte, kus erinevuse tingis asjaolu, et väljavõtte ja definitsioonikandidaat olid erinevates keeltes.

Hindamine oli range: juhul kui mudel oli leidnud väljavõttest info, mida tõlgendas poolikuna, ja liitis vastuses selle mõne samasuguse alguse või lõpuga tekstiga (5), loeti vastus valeks. Nii toimiti ka juhul, kui originaali vaadates sai mudeli üldistus kinnituse ja süsteemi kasutajale viga välja ei paistaks.

(5) Päring: ***aircraft arresting cable***

Väljavõtte: craft arresting system which spans the runway surface or flight deck landing area and is engaged by the aircraft arresting hook. [---]

Definitsioonikandidaat: **That portion of an aircraft** arresting system which spans the runway surface or flight deck landing area and is engaged by the aircraft arresting hook.

Eraldi vaadeldi juhte, kus definitsioonikandidaati ei saanud väljavõttega automaatselt kokku viia, kuna tulemuses esitatud allikaviide erines väljavõttes olevast. Selliseid näiteid oli 16 (1,2% 1320-st). Enamasti oli viide vale, sest pealkirjast oli osa eemaldatud. Näiteks dokumendi ATP-3.3.2.1(B) täpsustatud alapealkiri ei esinenud väljundis. Terminoloogi jaoks on viide küll piisavalt täpne, aga analüüsikriteeriume järgides loeti valeks. Ühel juhul oli vastuses leheküljenumbrina kasutatud infot, mis pärines väljavõttes leidunud jalusest. Kuigi originaaliga võrreldes võis ka taolist viidet pidada õigeks, loeti see valeks. Isegi kui viite väljakirjutus polnud ootuspärasel kujul, oli kõigi vigade puhul näha, et mudel oli lähtunud sisendandmetest. Kahel juhul oli viide siiski korrektne, ent ei läbinud automaatkontrolli mudeli vormiliste kõrvalekallete pärast, mida regulaaravaldis ei katnud. Kokku parandati õigeks 52 väljavõtte ja definitsioonikandidaadi erinevust ja 2 viidet, mis tähendab, et käsitsi kontrolli järel kattuvad kandidaat ja väljavõtte 92,5% juhtudest ning 98,9% viidetest on korrektsed.

RAG-süsteemi maksimumsuutlikkuse väljaselgitamine vajab põhjalikumat analüüsi, kuid süsteemi abiga on terminoloogidel juba praegu võimalus leida potentsiaalseid definitsioone piisavalt täpselt ja kordades tõhusamalt, samuti esitada allikaviiteid vajadusel ka leheküljetäpsusega. Võrreldes seni kasutusel oleva märksõnaotsinguga hulgast dokumentidest, tagab RAG-süsteem paindlikuma, kiirema ja süsteemsema lahenduse. Ometi pole see ideaalne. Lubamatuid kõrvalekaldeid võib tekkida dokumentide parsimisel, kus faili leheküljenumber ei pruugi ühtida dokumendi omaga. Teine probleemkoht on selete parsimine, mis võib väljavõtte sisu võrreldes dokumendi algversiooniga moonutada. Edaspidine arendus peab kindlasti neid kitsaskohti püüdma lahendada.

6. Tulevikusuunad

Kuigi kirjeldatav projekt pole veel lõppenud ning põhjalikum analüüs ja tulemuste hindamine leiavad aset 2024. aasta lõpus⁴, on juba praeguses arendusetapis võimalik nimetada suundi ja arenduskohti, mis võimaldavad suurendada prototüübi funktsionaalsust.

Tuleb arvestada, et TI-ga seotud süsteemid on kiires arengus ja isegi lähituleviku vaates pole võimalik prognoosida, millised aspektid mõjutavad RAG-süsteemide arengut ja arenduse järgmisi etappe. Projekti jooksul on lisandunud uuringuid, millest saab lähtuda nii pilootprojekti lõpufaasis kui ka edasiste tegevuste planeerimisel.

Üldisemas vaates võib näha kaht peamist arendussuunda. Esmatähtis on prototüübi universaalsus: süsteem peaks saama käepäraseks abivahendiks teistegi valdkondade terminitöös ja terminibaaside koostamisel. Järgmises etapis on plaanis kaasata külgnevaid valdkondi, esimene neist on sisekaitse. Teiseks suunaks on funktsioonide tõhustamine ja täiendamine, nt võiks süsteemi hakata katsetama eesti-, ent ka muukeelse ja enam kui kakskeelse sisendiga. Erinevatel erialadel on siin erinevad vajadused ja võimalused. Kui riigikaitstes on ajakohaseid eestikeelseid sisendandmeid pigem napilt, siis kindlasti on valdkondi, kus see nii ei ole. Küll aga on riigikaitse seisukohast omaette väljakutse vene keele lisamine, mille järele on tegelikkuses üha suurem vajadus. Samuti napib baasis prantsuskeelset teavet, mis on NATO teise ametliku keelena tähtis.

7. Kokkuvõte

Artikli eesmärk oli riigikaitseterminoloogia varal välja selgitada, kuivõrd sobib RAG-süsteem alusdokumentidest definitsioonide väljatoomiseks ja allikaviidete lisamiseks. Tegemist on ühe tööloiguga pilootprojektist, mille siht on riigikaitse terminitööle tuginedes välja töötada terminiarendust toetav, TI-l põhinev prototüüp.

RAG on nüüdisajal laialt kasutatav ja arendatav meetod, sest võimaldab SKM-idele edastada lisaandmeid, mille toel suudavad need anda faktitäpseid vastuseid. Kuna SKM-ide treenimiseks ei kasutata üldjuhul erialaspetsiifilisi andmeid, siis ei suuda mudelid vastata kitsaid valdkondi puudutavatele küsimustele. Seega on RAG-i abil võimalik mudelite vastuseid kohandada enda vajadustele vastavaks, samal ajal vähendades ebatäpse info hulka. Lisaks annab RAG-i rakendamine mudelile võime viidata allikatele, kust vastuses esitatud info pärineb. Ka annab RAG-süsteem võimaluse hiljem algandmetesse uusi dokumente lisada.

Esialgne analüüs näitas, et RAG-süsteemi täpsus alusandmetest definitsioonide eristamisel ja allikaviidete lisamisel on kõrge (vastavalt 89,6% ja 98,9%). Siiski on asjaolusid, mida peab RAG-süsteemi kasutamisel erialasõnastiku koostamisel arvesse võtma. Esiteks on oht, et dokumentide parsimisel andmed moonduvad. Teiseks võib tekkida olukord, et mudel esitab täpse definitsioonina üldistuse, mis põhineb mitmel erineval väljavõttel. Kolmandaks võivad viidetes esineda ebatäpsused, mis ei võimalda süsteemi kasutada viitamistööriistana, vaid pigem abivahendina alusandmetes lihtsamini orienteerumiseks.

⁴ Hindamise esialgsed tulemused viitavad, et süsteemi põhikomponendid toimivad. Samas on vajalik edasiarendus, et süsteemi pakutavad vastused oleksid sisukamad, täpsemad ning kasutajale läbipaistvamad.

Viidatud kirjandus

- Bogdanov, Dan; Etti, Paula; Kamm, Liina; Ostrak, Andre; Stomakhin, Fedor; Toomsalu, Maria; Valdma, Sandhra-Mirella; Veldre, Anto 2024. Tehisintellekti ja masinõppe tehnoloogia riskide ja nende leevendamise võimaluste uuring. RIA. <https://www.ria.ee/sites/default/files/documents/2024-03/Tehisintellekti-masinoppe-tehnoloogia-riskide-uuring-2024.pdf> (8.8.2024).
- Cram, Damien; Daille, Béatrice 2016. Terminology extraction with term variant detection. – Sameer Pradhan, Marianna Apidianaki (Eds.), *Proceedings of ACL-2016 System Demonstrations*. Berlin: Association for Computational Linguistics, 13–18. <https://doi.org/10.18653/v1/P16-4003>
- Cuconasu, Florin; Trappolini, Giovanni; Siciliano, Federico; Filice, Simone; Campagnano, Cesare; Maarek, Yoelle; Tonello, Nicola; Silvestri, Fabrizio 2024. The power of noise: Redefining retrieval for RAG systems. – *Proceedings of ACM SIGIR'24*. New York: Association for Computational Linguistics, 719–729. <https://doi.org/10.1145/3626772.3657834>
- de Schryver, Gilles-Maurice 2023. Generative AI and lexicography: The current state of the art using ChatGPT. – *International Journal of Lexicography*, 36 (4), 355–387. <https://doi.org/10.1093/ijl/ecado21>
- Fatehkia, Masoomali; Lucas, Ji Kim; Chawla, Sanjay 2024. T-RAG: Lessons from the LLM trenches. – arXiv:2402.07483. <https://doi.org/10.48550/arXiv.2402.07483>
- Gao, Yunfan; Xiong, Yun; Gao, Xinyu; Jia, Kangxiang; Pan, Jinliu; Bi, Yuxi; Dai, Yi; Sun, Jiawei; Wang, Meng; Wang, Haofen 2024. Retrieval-augmented generation for large language models: A survey. – arXiv:2312.10997. <https://doi.org/10.48550/arXiv.2312.10997>
- Gross, Daniel P.; Sampat, Bhaven N 2020. Inventing the endless frontier: The effects of the World War II research effort on post-war innovation. – *Harvard Business School Working Paper*, 20–126. <https://doi.org/10.2139/ssrn.3623115>
- Hendrikson, Reet 2014. Tõhususe tahud sõjandusterminoloogia näitel: ülediferentsimine või erinev tunnetusväärtus? – *Eesti Rakenduslingvistika Ühingu aastaraamat*, 10, 107–123. <https://doi.org/10.5128/ERYa10.07>
- Hendrikson, Reet 2018. Kas sõjasõna sünnib sõtta? Erialakeele tõhusus sõjandusterminoloogia näitel. *Dissertationes philologiae estonicae Universitatis Tartuensis*, 42. Tartu: Tartu Ülikooli Kirjastus.
- Hendrikson, Reet 2023. Riigikaitse terminiarendus – kerkne kriiside kiuste! – *Õiguskeel*, 4. Kaitseministeeriumi valitsemisala teadus- ja innovatsioonipoliitika. Kaitseministeerium, 2022. https://kaitseministeerium.ee/sites/default/files/kaitseministeeriumi_valitsemisala_teadus-_ja_innovatsioonipoliitika.pdf (14.10.2024).
- Kilgarriif, Adam; Baisa, Vít; Bušta, Jan; Jakubíček, Miloš; Kovář, Vojtěch; Michelfeit, Jan; Rychlý, Pavel; Suchomel, Vít 2014. The Sketch Engine: ten years on. – *Lexicography*, 1 (1), 7–36. <https://doi.org/10.1007/s40607-014-0009-9>
- Koppel, Heidi 2024. Allikapõhiste genereerimissüsteemide loodud vastuste kvaliteedi hindamine. BA. Tallinna Tehnikaülikool, infotehnoloogia teaduskond.
- Koppel, Kristina; Kallas, Jelena 2022. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. – *Eesti Rakenduslingvistika Ühingu aastaraamat*, 18, 207–228. <https://doi.org/10.5128/ERYa18.12>
- Lamothe, Maxime; Guéhéneuc, Yann-Gaël; Shang, Weiwei 2021. A systematic review of API evolution literature. – *ACM Computing Surveys*, 54 (8), A171, 1–36. <https://doi.org/10.1145/3470133>
- Lewis, Patrick; Perez, Ethan; Piktus, Aleksandra; Petroni, Fabio; Karpukhin, Vladimir; Goyal, Naman; Küttler, Heinrich; Lewis, Mike; Yih, Wen-tau; Rocktäschel, Tim; Riedel, Sebastian; Kiela, Douwe 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. – *Advances in Neural Information Processing Systems*, 33, 9459–9474.

- Ma, Xinbei; Gong, Yeyun; He, Pengcheng; Zhao, Hai; Duan, Nan 2023. Query rewriting in retrieval-augmented large language models. – Houda Bouamor, Juan Pino, Kalika Bali (Eds.), *Proceedings of the 2023 Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, 5303–5315. <https://doi.org/10.18653/v1/2023.emnlp-main.322>
- Massion, François 2024. Terminology in the age of AI: The transformation of terminology theory and practice. – *Journal of Translation Studies*, 4 (1), 67–94. <https://doi.org/10.3726/JTSO12024.04>
- McKean, Erin; Fitzgerald, Will 2024. The ROI of AI in lexicography. – *Lexicography*, 11 (1), 7–27. <https://doi.org/10.1558/lexi.27569>
- Mills, M. Anthony; Mills, Mark P 2014. The invention of the war machine: Science, technology, and the First World War. – *The New Atlantis*, 42, 3–23.
- Rees, Geraint Paul; Lew, Robert 2024. The effectiveness of OpenAI GPT-generated definitions versus definitions from an English learners' dictionary in a lexically orientated reading task. – *International Journal of Lexicography*, 37 (1), 50–74. <https://doi.org/10.1093/ijl/ecado30>
- Reineke, Detlef 2023. *Terminologearbeit mit ChatGPT & Co.* – edition, 1, 25–28.
- San Martín, Antonio 2024. What generative artificial intelligence means for terminological definitions. – *Proceedings of the 3rd International Conference on Multilingual Digital Terminology Today* 2024. <https://ceur-ws.org/Vol-3703/paper1.pdf> (19.3.2025).
- Strickland, Jonathan 2023. Do wars drive technological advancement? – *HowStuffWorks*, 9.6. <https://science.howstuffworks.com/war-drive-technological-advancement.htm> (14.10.2024).
- Zhang, Yue; Li, Yafu; Cui, Leyang; Cai, Deng; Liu, Lemao; Fu, Tingchen; Huang, Xinting; Zhao, Enbo; Zhang, Yu; Chen, Yulong; Wang, Longyue; Luu, Anh Tuan; Bi, Wei; Shi, Freda 2023. Siren's song in the AI ocean: A survey on hallucination in large language models. – *arXiv:2309.01219*. <https://doi.org/10.48550/arXiv.2309.01219>
- Uchida, Satoru 2024. Using early LLMs for corpus linguistics: Examining ChatGPT's potential and limitations. – *Applied Corpus Linguistics*, 4 (1), 1–9. <https://doi.org/10.1016/j.acorp.2024.100089>
- Wang, Liang; Yang, Nan; Huang, Xiaolong; Yang, Linjun; Majumder, Rangan; Wei, Furu 2024. Multilingual e5 text embeddings: A technical report. – *arXiv:2402.05672*. <https://doi.org/10.48550/arXiv.2402.05672>
- Wright, Quincy 1941. Inventions and War. – *The Scientific Monthly*, 53 (6), 526–541.

Võrguviited

- AKIT = Andmekaitse ja infoturbe portaal. <https://akit.cyber.ee/term/17003-retrieval-augmented-generation> (24.09.2024).
- Collins API. <https://www.collinsdictionary.com/collins-api> (8.8.2024).
- Collins English Dictionary. <https://www.collinsdictionary.com/dictionary/english> (15.10.2024).
- GPT-4o. <https://openai.com/index/hello-gpt-4o/> (15.10.2024).
- GPT-4o mini. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/> (15.10.2024).
- IATE. <https://iate.europa.eu/home> (15.10.2024).
- IATE2 Search API. <https://documenter.getpostman.com/view/4028985/RztoMTwn> (8.8.2024).
- Merriam-Webster Dictionary. <https://www.merriam-webster.com/> (15.10.2024).
- Merriam-Webster Dictionary API. <https://dictionaryapi.com/application/products/index> (8.8.2024).

Militerm: Riigikaitseterminite andmebaas. Sõnaveeb. <https://sonaveeb.ee/ds/mil> (10.10.2024).

Multilingual-e5-Large. <https://huggingface.co/intfloat/multilingual-e5-large> (17.10.2024).

NATO AI-strateegia = Summary of NATO's Revised Artificial Intelligence (AI) Strategy. https://www.nato.int/cps/en/natohq/official_texts_227237.htm (8.8.2024).

OpenAI: Models. <https://platform.openai.com/docs/models> (15.10.2024).

TermSuite. <https://termsuite.github.io/> (24.9.2024).

Sandra Eiche (Eesti Keele Instituut) on keeletehnoloog, kelle huviks on keeletehnoloogiate praktilised rakendused.
Roosikrantsi 6, 10119 Tallinn, Estonia
sandra.eiche@eki.ee

Reet Hendrikson (Kaitseväe Akadeemia) on Kaitseväe juhtivterminoloog, riigikaitseterminoloogia töörühma juht ja kuulub Eesti terminitöö juhtkomiteesse. Peamised uurimissuunad on erialakeel, selle tõhusus ja kujundlikkus.
Riia 12, 51013 Tartu, Estonia
Reet.Hendrikson@mil.ee

Eleri Aedmaa (Eesti Keele Instituut) on keeletehnoloog, kes on keskendunud suurtele keelemudelitele. Varem uurinud statistiliste meetodite rakendamist keeleteaduses.
Roosikrantsi 6, 10119 Tallinn, Estonia
eleri.aedmaa@eki.ee

Esta Prangel (Eesti Keele Instituut) on Ekilexi tootejuht, varem uurinud soome-ugri keeli.
Roosikrantsi 6, 10119 Tallinn, Estonia
esta.prangel@eki.ee

Mari-Liis Tikerperi (Kaitseväe Akadeemia, Tartu Ülikool) on riigikaitseterminoloogia töörühma spetsialist, teadustöö fookuses on üldhariduskoolide kommunikatsioonijuhtimine.
Riia 12, 51013 Tartu, Estonia / Ülikooli 18, 50090 Tartu, Estonia
Mari-Liis.Tikerperi@mil.ee
Mari-Liis.Tikerperi@ut.ee

APPLYING ARTIFICIAL INTELLIGENCE IN SPECIALIZED LANGUAGE DEVELOPMENT: THE EXAMPLE OF ESTONIAN NATIONAL DEFENSE TERMINOLOGY

**Sandra Eiche¹, Reet Hendrikson²,
Eleri Aedmaa¹, Esta Prangel¹, Mari-Liis Tikerperi^{2,3}**

Institute of the Estonian Language¹, Estonian Military Academy², University of Tartu³

This article explores the application of Retrieval-Augmented Generation (RAG) for extracting definitions and source references in defense terminology. By integrating external data during generation, RAG enhances large language models, reducing hallucinations and improving accuracy. This capability is critical in fields like defense terminology, where precise definitions and reliable references are essential.

The study found that RAG accurately extracts definitions and references from foundational documents, achieving a precision of 89.6% for definitions and 98.9% for references. However, challenges exist, particularly in parsing documents like PDFs, which can lead to occasional inaccuracies. These issues highlight the importance of terminologists for ensuring data quality, as RAG cannot fully replace human oversight.

Despite these limitations, RAG holds promise as a tool to support terminology work by efficiently retrieving relevant data and assisting in dictionary compilation. It also offers potential for enhancing other stages of terminology development, such as suggesting synonyms and identifying issues in source texts. Further improvements, such as refining document parsing and reference validation, could enhance RAG's reliability, making it a valuable tool for terminologists.

Keywords: retrieval-augmented generation (RAG), LLM, specialized language, terminology, definition, concept, term, defense, English