TALLINNA TEHNIKAÜLIKOOL
Infotehnoloogia teaduskond

Mari-Anna Meimer 222854IAIB
Loore Lehtmets 222963IAIB

AJALOOLISTE EESTIKEELSETE OCR TEKSTIDE JÄRELTÖÖTLUSE JA HINDAMISE AUTOMATISEERIMINE EESTI RAHVUSRAAMATUKOGU JAOKS

Bakalaureusetöö

Juhendaja: Priit Järv
PhD

Tallinn 2025

Autorideklaratsioon

Kinnitame, et oleme koostanud antud lõputöö iseseisvalt ning seda ei ole kellegi teise poolt varem kaitsmisele esitatud. Kõik töö koostamisel kasutatud teiste autorite tööd, olulised seisukohad, kirjandusallikatest ja mujalt pärinevad andmed on töös viidatud.

Autorid: Mari-Anna Meimer, Loore Lehtmets

16.05.2025

Annotatsioon

Lõputöö tulemusena valmis Eesti Rahvusraamatukogule kahte eesmärki täitvad keelemudelid: mudelid, mis parandavad OCR tekstis vigu ja mudelid, mis hindavad OCR-i parandava mudeli poolt parandatud teksti ja algse OCR teksti kvaliteeti. Mudelid on mõeldud kasutamiseks rahvusraamatukogu töötajatele, et nende tööd ajalooliste tekstide digiteerimisel lihtsustada ja kiirendada.

Lõputöö eesmärgiks on, et koostöös suudavad mudelid parandada valitud andmestiku üldkvaliteeti. Paranduste mudelilt ootame, et vähemalt poolte testnäidete kvaliteet paraneks ja hindamismudelit, et vähemalt kõige halvemad ning kõige paremad näited oleksid õigesti tuvastatud.

Töö tulemusena valminud paranduste mudelid küll parandavad või jätavad samaks üle poolte tekstide CER-i, aga kõige halvema kvaliteediga parandustest tingituna on keskmine muutus üle testandmestiku siiski väikeses miinuses. CER-i hindav mudel on oodatust parem, sest mediaanerinevus tegeliku ja ennustatud CER-i vahel on vaid 2%. Paranduste hindamise mudeli täpsus on üldjoontes rahuldav, sest ligikaudu 62% parandusi klassifitseeritakse õigesti (parem/halvem kui OCR tekst), aga vähemalt halvimad ja parimad parandused klassifitseeritakse enamjaolt õigesti. Mudelite koostöö tulemusena keskmine CER paraneb üle testandmestiku.

Lõputöö on kirjutatud eesti keeles ning sisaldab teksti 88 leheküljel, 9 peatükki, 20 joonist, 17 tabelit.

Abstract

Automation of Post-Processing and Evaluation of Historical Estonian OCR Texts for the National Library of Estonia

As a result of the thesis, two types of language models will be developed for the National Library of Estonia: models that correct errors in OCR texts and models that evaluate the quality of the texts corrected by the correction model, and the original OCR text. The models are intended for use by the National Library staff to facilitate and speed up their work in digitizing historical texts.

The goal of the thesis is that the models can improve the overall quality of the selected dataset through collaboration. We expect the correction model to improve the quality of at least half of the test examples, and the evaluation model to correctly identify at least the worst and best examples.

The correction models improve or maintain the CER for more than half of the texts, but due to the worst corrections' high CER, the average change across the test dataset is still slightly negative. The CER estimation model performs better than expected, as the median difference between the actual and predicted CER is only 2%. The accuracy of the correction estimation model is generally satisfactory, as about 62% of corrections are classified correctly (better/worse than the OCR text), but at least the worst and best corrections are classified correctly for the most part. As a result of the cooperation of the models, the average CER improves across the test dataset.

The thesis is written in Estonian and is 88 pages long, including 9 chapters, 20 figures and 17 tables.

Lühendite ja mõistete sõnastik

API	Rakendusliides (*Application Programming Interface*)
CPU	Keskseade (*Central Processing Unit*)
OCR	Optiline märgituvastus (*Optical Character Recognition*)
SKM	Suur keelemudel (*Large Language Model (LLM)*)
FT	Peenhäälestamine (*fine-tuning*)
CER	Keskmine vigade arv tähemärgi kohta (*Character Error Rate*)
WER	Keskmine vigade arv sõna kohta (*Word Error Rate*)
DPO	Otsene eelistuste suunamine (*Direct Preference Optimization*)
RLHF	Inimtagasisidel põhinev stiimulõpe (*Reinforcement Learning from Human Feedback*)
GRPO	Juhendatud preemiasüsteemi optimeerimine (*Guided Reward Policy Optimization*)
Reward Model	Preemiamudel
Seq2Seq	Järjestus-järjestusmudelid (*Sequence-to-Sequence model*)
RaRa	Eesti Rahvusraamatukogu
DIGAR	Eesti Rahvusraamatukogu digitaalarhiiv
Viip	Keelemudelilt tavakeeles küsitud küsimus või päring (*prompt*)
Näitetu viip	Keelemudelilt tavakeeles küsitud küsimus või päring ilma ühegi näiteta (*zero-shot prompt*)
Ainunäitega viip	Keelemudelilt tavakeeles küsitud küsimus või päring ühe näitega (*one-shot prompt*)
Mitmiknäitega viip	Keelemudelilt tavakeeles küsitud küsimus või päring mitme näitega (*few-shot prompt*)
GT	Inimparandatud tekst ehk tõeandmed (*ground truth*)

Sisukord

Jooniste loetelu

Tabelite loetelu

1. Sissejuhatus

Oluline samm kultuuripärandi pikaajaliseks ja turvaliseks säilitamiseks on ajalooliste dokumentide digiteerimine. Selle jaoks kasutab Eesti Rahvusraamatukogu (edaspidi RaRa) erinevaid optilise märgituvastuse (edaspidi OCR) mudeleid, mis võivad digiteerimisel teha vigu. Aja jooksul on küll RaRa OCR mudelite sooritus paranenud, aga arhiivis on veel palju tekste, mis on vanemate mudelitega tuvastatud ja seetõttu sisaldavad endas palju vigu. Samuti mõjutab OCR kvaliteeti näiteks originaaldokumendi kvaliteet (Lisa 2).

Digitaalarhiivi (edaspidi DIGAR) kasutajatel on olnud võimalus vigaseid tekste käsitsi parandada, kuid RaRa on hakanud otsima võimalusi seda tööd automatiseerida. RaRa andmekogus on üle 800 000 inimkontrollitud rea, millesse on panustanud üle 500 inimese [1]. Parandustega seoses esineb aga mitmeid probleeme. Esiteks on väga vigaste tekstide parandamine aeganõudev tegevus ning kuna parandusi on teostatud ühisloome meetodil, ei saa neid täieliku tõena võtta.

Teine probleem seisneb tekstiparanduste kvaliteedi hindamises. Praegu ei ole RaRal täielikku ülevaadet, millise OCR kvaliteediga mingi konkreetne tekst on. Varem on teostanud valitud digiteeritud tekstide hindamist ChatGPT-4o mini mudeliga RaRa andmeteadlane Krister Kruusmaa. Paljude keelemudelite, sealhulgas ka ChatGPT, kitsaskohaks on aga eestikeelsete tekstide töötlemine ning grammatikareeglitest kinnipidamine. Eriti tekitavad probleeme vanema keelekasutusega tekstid, kus näiteks “v” asemel on kasutusel “w” või lauseehitus erineb tänapäeval levinud kirjapildist. Samuti võib kommertsmudelite kasutamine osutuda suure hulga töötlemist vajavate tekstide puhul rahaliselt kulukaks tegevuseks ning paljud RaRa tekstid on piiratud autoriõigustega, mistõttu ei tohigi neid üldse kommertsmudelitele parandamiseks anda.

Lõputöö tulemusena valmivad mudelid täidavad kahte erinevat ülesannet: OCR tekstides vigade parandamine ja esimese mudeli poolt parandatud tekstide ning algsete OCR tekstide kvaliteedi hindamine. Mudelid on mõeldud RaRa töötajate tööprotsessi kiirendamiseks ja lihtsustamiseks.

Kasutame tekstide parandamiseks ja hindamiseks keelemudelit, mida treenime enamjaolt ajalooliste ajalehtede ja ajakirjade, kohati ka ajalooliste ilukirjandustekstide peal. Valimi suurendamiseks genereerime tõenäosuste põhjal vigaseid sünteetilisi tekste.

2. Taust

2.1 RaRa DIGAR

Euroopa Regionaalarengu Fondi rahastatud "Kultuuripärandi digiteerimine 2018-2023" projekti käigus digiteeriti 1,5 miljonit lehekülge tekste, millest suurema osa moodustab nõukogudeaegne ajakirjandus. Esindatud on 1703 ajalehte, 650 ajakirja ning 550 jätkväljaannet [2].

Esimesed portaalis olevad ajalehed on välja antud juba 1811. aastal, ajakirjad ja jätkväljaanded on portaalis aastast 2017. Portaali täiendatakse artiklitega iga päev ning hetkel on saadaval 689 589 väljaannet, 6 557 786 lehekülge ja 18 635 996 artiklit [2].

DIGAR-is on iga digiteeritud väljaande juures vastav PDF-dokument, millele on rakendatud OCR-tehnoloogiat. Digiteeritud tekst on tavaliselt tuvastatud veergude või mõnikord ka artiklite kaupa. Mõne väljaande puhul on aga lehekülje kujundus väga tihe, mistõttu on tekst tuvastatud järjestikku üle kogu lehekülje, vasakult paremale. Näiteid DIGAR-is olevatest digiteeritud väljaannetest on leitavad lisast 2.

Tekste saavad parandada ainult registreerunud kasutajad ning ainult selliseid väljaandeid, mis on ilmunud enne 1944. aastat [3]. DIGAR-i üks probleem on veel see, et ei omata informatsiooni selle kohta, milline OCR mudel konkreetset teksti töötles. Kuna mudeleid on pidevalt uuendatud aastate jooksul, on sellest ka kvaliteet paranenud, kuid kahjuks DIGAR-is ei ole võimalik öelda enne teksti nägemist, milline selle eeldatav OCR kvaliteet olla võiks ning sellest tingituna kui ajamahukas võiks olla selle parandamine.

2.2 RaRa varasemad OCR tegevused

Kuigi RaRa OCR mudelite sooritus on aja jooksul paranenud, siis OCR järeltöötlusele pühendatud mudelit soovivad nad osaliselt ka sellepärast, et kõiki kehva OCR kvaliteediga tekste ei peaks uuesti transkribeerima. Samuti aitab hindamise mudel tuvastada tekstid, mille puhul OCR kvaliteet on nii halb, et need tuleb uuesti töödelda ja automaatne parandamine võimalik ei ole.

2024. aastal hindas RaRa andmeteadlane Krister Kruusmaa eksperimendi korras valitud tekstide OCR kvaliteeti DIGAR-is, kasutades selleks ChatGPT-4o mini mudelit [1, 4].

Mudelile anti sisendiks hindamisskaala (hinded skaalal 1-5 koos kirjeldustega, koostatud RaRa poolt) ning tekst, mida hinnata. Hindamisskaala koos ChatGPT-le antud viibaga on leitav lisast 3.

Samuti proovis Kruusmaa ChatGPT-3.5 peenhäälestamist 300, 700 ja 1800 näitega ning testis peenhäälestatud mudeleid 11 näitega. Kui 300 ja 700 näite puhul üldpilt paranes, siis 1800 näitega läksid tulemused hoopis halvemaks [5]. Eksperimendi tulemuste analüüsimise ning edasiarenduse eesmärgil kasutame Kruusmaa koostatud andmestikku ka meie.

2.2.1 Ühisloome probleemid

Senimaani on suurima panuse andnud OCR tekstide parandamisse DIGAR-i kasutajad, keda on viimaste andmete järgi üle 500 ning kelle panusega on ühisloome meetodil parandatud üle 800 000 rea [1]. Parandustega esineb aga mitmeid probleeme.

Esiteks on käsitsi parandamine aeganõudev tegevus, sest paljude OCR tekstide kvaliteet on väga madal ning parandamiseks peab skaneeritud ajalehe pealt vastava koha üles otsima. Teine, ning meie töö kontekstis palju suurem probleem on see, et paranduste kvaliteet ei ole kontrollitav. Kuna parandada võib igaüks ja mistahes mahus, siis ei saa veenduda paranduste korrektsuses ja terviklikkuses.

Ka meie andmestikus on hulganisti tekste, mille parandamine on selgelt pooleli jäetud või kus parandajal on sisse jäänud kirjavigu. Parandajatel ei ole ka kindlat reeglistikku mida järgida ehk tekstid on parandatud erinevalt - näiteks on mõned parandajad ajalehes poolitatud sõnad kirjutanud kokku, mõned lahku, mõned sidekriipsuga, mõned võrdusmärgiga jne.

2.3 Kommertsmudelid ja nende kasutamine

Põhjuseid, miks ajalooliste tekstide digiteerimisel kommertsmudelite kasutamist vältida, on mitmeid. Näiteks ChatGPT-4o mudeli API kasutamiseks miljon *token*-it maksab sisendil kaks ja pool dollarit ning väljundil kümme dollarit. Kui tegelikult on aga üks *token* umbes neli tähemärki, siis treening- ja testandmestiku suurusest olenevalt võivad lõppsummad olla sadade või isegi tuhandete eurode ringis. Samuti ei ole lubatud saata kommertsmudelile piiratud õigustega tekste, mida RaRa DIGAR-is esineb. Lisaks on meile teadaolevalt küll keelemudeleid, mis toetavad eesti keelt, aga nende keeleoskuse tase on võrreldes inglise keelega veel nõrk.

2025. aasta jaanuaris andis Hiina tehnoloogiaettevõte DeepSeek välja vabavaralise DeepSeek R1 vestlusroboti [6]. DeepSeeki eelis paljude teiste kommertsmudelite ees on see, et 2025. aasta märtsi seisuga on Hugging Face'i leheküljel DeepSeek'i profiili all saadaval 69 erinevat mudelit, mida saab igaüks soovi korral vastavalt enda vajadustele kasutada ning ei tekita probleeme seoses autoriõigustega piiratud tekstidega, küll aga on selleks vaja suurt arvutusvõimekust[1].

Samuti pakub DeepSeek sarnaselt OpenAI-ga API päringute võimalust, kuid võrreldes näiteks eelmainitud ChatGPT-4o mudeliga on ühe miljoni *token*-i hind pea kümme korda madalam (sisendil 0.27 dollarit ning väljundil 1.10 dollarit) [7]. Kuigi hind on palju odavam, siis ühtlasi on DeepSeek päringud ka kordades aeglasemad ning suurte andmemahtude puhul on see probleemne. Lisaks just R1 mudeli puhul, kus mudel lisaks väljundile tagastab ka sellele eelnenud mõttekäigu, kujuneb hind ikkagi kallimaks, kuna pikad mõttekäigud tõstavad hinda märkimisväärselt. Lisaks on olemas veel vestlusmudel DeepSeek V3. DeepSeek mudelite eesti keele oskus on veel võrreldes inglise ja hiina keelega nõrk, kuna mudeleid ei ole sihilikult eesti keelega treenitud.

2.4 Olukord mujal maailmas

Kuigi OCR tehnoloogiad on erineval kujul eksisteerinud umbes 100 aastat ning tänapäevased tehnoloogiad võivad teatud kasutusjuhtudel saavutada isegi üle 99% täpsuse, siis paljudes praktilistes rakendustes jäävad need siiani inimsilmale alla [8, 9]. Üheks selliseks rakenduseks on ajaloolised tekstid, kus võib esineda mitmeid segavaid faktoreid, näiteks laialivalgunud tint, vanaaegsed ja sageli keerukad fondid ning lihtsalt aja jooksul tekkinud kulumisjäljed. Lisaks võivad ajalehtede kujundus ja tekstipaigutus olenevalt väljaandest suuresti erineda, mistõttu tuvastatud tekst jookseb kohati terve lehe ulatuses vasakult paremale, mitte tulpade kaupa. Kuna vanemad ajalehed on enamjaolt mustvalged, siis mõjutab lehe kujundus OCR-i kvaliteeti veelgi enam, sest pilte ja teksti on üksteisest raskem eristada. Seetõttu võib leida üpriski erinevate lõpptulemustega järeltöötluse eksperimente.

Lihtsamate ja uuemate tekstide puhul ollakse tuleviku suhtes optimistlikud, kuid ajalooliste tekstide töötlemisel ei ole märkimisväärseid tulemusi saavutatud isegi inglise keele puhul [10, 11]. Erinevate eeltöötlusvahenditega on mudelite töökvaliteeti võimalik küll parandada, ent väikekeelte jaoks on OCR järeltöötlus siiani suur väljakutse. 2025 aasta märtsis ilmunud artiklis proovisid Turu Ülikooli teadlased treenida erinevaid mudeleid nii inglisekui ka soomekeelsete ajalooliste tekstide peal. Positiivse tulemuse saavutasid inglise keele

[1] Hugging Face DeepSeek `https://huggingface.co/deepseek-ai`

peal seitsmest mudelist kuus, samal ajal kui soomekeelsete tekstide peal suutis tulemusi parandada samade mudelite hulgast ainult üks [12]. Meile teadaolevalt ei ole keegi peale Kruusmaa veel eesti keele peal OCR järeltöötlust teha proovinud.

Samuti on meile teadaolevalt Kruusmaa Eestis ka ainuke, kes on OCR tekstide kvaliteedi hindamist teinud. Ka mujal maailmas on OCR tekstide kvaliteedi hindamist tehtud vähe ning peamiselt on keskendutud vigade parandamisele. Siiski, 2022. aastal avaldatud artiklist (Booth jt.) selgub, et ajastu ning temaatika poolest kitsa tekstikorpuse peal on võimalik treenida mudel, mis on edukas kvaliteetsete OCR tekstide välja filtreerimises [13]. Samuti on 2022. aastal Luksemburgi raamatukogu uurinud võimalikke lahendusi OCR kvaliteedi hindamiseks ning võttes osaliselt arvesse ka uue OCR mudeli võimekust, kuna nemad soovisid seda kasutada peamiselt eesmärgiga, et leida tekste, mida oleks vaja ja võimalik uuesti OCR tehnoloogia abil töödelda. Nemad ei kasutanud hindamiseks keelemudelit, vaid hoopis traditsioonilisi masinõppe tehnikaid ning hindasid enda tulemused edukaks [14].

3. Metoodika

Paranduste mudeli treenimiseks katsetasime kahte erinevat metoodikat: peenhääletamist ja eelistuste suunamist. Hindamise mudeli puhul katsetasime peenhäälestamist ja preemiamudeli treenimist. Tulemuste valideerimiseks kasutame paranduste mudelite puhul CER ja WER parameetreid ning võrdleme tulemusi kommertsmudelite tasemega. Hindamise puhul võrdleme tulemusi samuti kommertsmudelite saavutatud tulemustega, kuid hindame ennustatud hinnete täpsust. Andmestikud sisaldavad kahte tüüpi tekste: Kruusmaa kogutud tekstid RaRa DIGAR-ist ning tekstid, millele on lisatud sünteetilisi vigu.

Suurte keelemudelite (edaspidi SKM) treenimisel ja käitamisel on oluline kasutada läbimõeldud viipasid ning oma valikuid tutvustame peatükis 3.1. Seejärel toome põhjenduse SKM-ide kasutamiseks ning tutvustame erinevaid meetodeid, mida kasutame, et neid meie ülesande jaoks efektiivselt käitada (3.2). Täpsemalt räägime konkreetse mudeli valikust (3.2.1), andmestiku suurendamisest sünteetiliste andmetega (3.2.2), ning tulemuste valideerimisest (3.2.3). Samuti tutvustame erinevaid hindamismudeli treenimistaktikaid (3.3). Samuti räägime hindamismudeli peatükis mudeli valikust (3.3.2) ja tulemuste valideerimisest (3.3.3). Viimaseks põhjendame tehnoloogiavalikuid (3.4), täpsemalt programmeerimiskeelt (3.4.1), keskkondi (3.4.2) ning analüüsimeetodeid peatükis 3.4.3.

3.1 Viipade läbimõeldud kavandamine

Viipade läbimõeldud kavandamisel (*prompt engineering*) on oluline roll suurte keelemudelite vastuste kujundamisel. Viipade koostamiseks on erinevaid tehnikaid, millest populaarseim on näitetu viip (*zero-shot prompt*). Sellise viiba puhul ei anna kasutaja mudelile ette näidet ning mudel kasutab vastuse andmiseks oma üldistamisoskust. Paljudes olukordades aitab aga kvaliteetsema vastuse saamiseks ainunäitega viiba (*one-shot prompt*) või mitmiknäitega viiba (*few-shot prompt*) kasutamine, et anda mudelile rohkem konteksti ja ülevaade oodatavast väljundist [15].

Lõputöö raames kasutame näitetut viipa, sest andmestikus sisalduvate tekstide kvaliteet, sisu, ajastu ja pikkus on väga varieeruvad. Seetõttu mudelile üksikute näidete etteandmine võib seda rohkem segadusse ajada kui aidata. Alguses proovisime üksikute näidete peal samuti ainu- ja mitmiknäitega viipa, kuid kuna tulemused ei paranenud ning kohati isegi halvenesid, siis kinnitas see veelgi esmast eeldust.

3.2 OCR tekstide järeltöötlus

Paranduste mudeli loomiseks peenhäälestame SKM-e. Enne SKM-ide populariseerumist ja laiemale avalikkusele kättesaamist oli OCR tekstides vigade parandamiseks kasutusel palju erinevaid meetodeid, näiteks: ühisloomemeetodil inimparandused, vigastele sõnadele sõnastikust kõige tõenäolisema vaste leidmine ning närvivõrkude või järjestusest-järjestusse (Seq2Seq ehk Sequence-to-Sequence model) mudelite kasutamine [16]. SKM-id eristuvad aga eelmainitud meetoditest oma üldistusvõime ning kontekstitundlikkuse poolest. See võimaldab mudeleid treenida laiemate tekstikorpuste peal ning ühel mudelil on seeläbi suurem potentsiaal hästi parandada laiemat hulka eriilmelisi tekste.

Näiteks võiksime enda töö kontekstis kasutada RaRa andmestiku, et iga tähe jaoks arvutada tõenäosused, millega see tekstis ekslikult asendatakse. Seeläbi saaksime tõenäosuste abil pöördprojekteerida vigastesse tekstidesse kõige tõenäosuslikumad parandused. Selle lähenemise juures on aga üks suur probleem: andmestikus on ajastust ja piirkonnast sõltuvalt väga erineva keelekasutusega tekste. Seetõttu ei saa lihtsalt eeldada, et iga "v" asemele käib "w" või iga "f"-i asemel käib tegelikult "s". Selle tõttu on kontekstil meie ülesande puhul väga oluline roll.

Lisaks SKM-ide peenhäälestamisele proovime treenimisel eelistuste suunamist (*preference alignment*). Eelistuste suunamisel antakse mudelile treeningprotsessi käigus ette kolm väärtust: viip, eelistatud väljund ning tagasilükatud väljund [17]. Selle abil peaks mudel ära tundma mustrid, mida kasutaja eelistab ning treenimise tulemusena andma tagasi kasutajale sobivamaid vastuseid. Meile teadaolevalt ei ole keegi seda meetodit veel OCR järeltöötlusel kasutanud ning loodame avastada uue võimaluse paranduste kvaliteedi tõstmiseks eelistuste suunamise näol.

Eelistuste suunamiseks kasutame tehnikat nimega DPO (*Direct Preference Optimization*) [18]. Selle eelis teiste eelistuse suunamise meetodite ees on peamiselt lihtsus, sest piisab vaid andmestiku eelnevalt mainitud kujule viimisest ning SKM-ist. Teisalt näiteks RLHF (*Reinforcement Learning from Human Feedback*) ja GRPO (*Guided Reward Policy Optimization*), mille lõppeesmärk on DPO-ga sama, vajavad eelnevalt eraldi *reward* mudeli ehk preemiamudeli treenimist. Preemiamudel peaks treenimise tulemusena õppima ära tundma eelistatud väjundeid ning selle abil suunama SKM-i treeningprotsessi [19, 20]. Samuti oleks vaja inimeste tagasisidet ja eelistusi, mis on konkreetse ülesande raames ajamahukas tegevus, kuna tekstide lugemine ja võrdlemine omavahel on keeruline. Seetõttu on kiireim lahendus DPO näol meie jaoks eelistatum valik.

3.2.1 Keelemudeli valik

Peenhäälestamiseks kasutame eestikeelsete andmete peal treenitud Llama-2 vestlusrobotit, mille avaldasid Tartu Ülikooli TartuNLP grupi teadlased 2024. aasta juunis [21]. Mudeli nimi on Llammas ehk eestindatud versioon Llamast. Kõige olulisem on mudelivaliku juures leida juba eesti keelt toetav variant, et peenhäälestamisel ei peaks alustama mudelile keele õpetamisest, vaid saaks kohe alustada vigade parandamisest. Lisaks Llammasele on eestikeelsete mudelite hulgas valikus samuti TartuNLP poolt loodud EstBERT mudel, aga kuna see avaldati juba 2020. aastal, siis kasutame värskemat lahendust [22].

Llammase alusmudeliks on tehnoloogiaettevõtte Meta poolt arendatud Llama-2-7b-hf ehk Llama-2 mudel, millel on 7 miljardit parameetrit (nime lõpus olev hf tähistab *Hugging Face Transformers format*) [23]. Esimese peenhäälestusringi tulemusena valmis Llammas-base mudel ehk Llammase baasmudel [24]. Baasmudeleid saadakse treenides mudeleid suurte tekstihulkade peal, et õpetada neile erinevaid keeli ja teemasid, aga need ei oska veel järgida juhiseid [25]. Llammase baasmudeli peenhäälestamise tulemusena valmis omakorda kaks Llammase versiooni: eelmainitud Llammase vestlusrobot ning Llammas kirjavigade parandamiseks [26]. Treenime omakorda kõiki kolme Llammase versiooni, et leida meie töö kontekstis kõige sobivam.

DeepSeek mudelite treenimine oleks samuti võimalik, kuid selleks, et treenida samu mudeleid, mida läbi API või veebiversiooni kasutada saab, on vaja suurt arvutusvõimekust. Samadest mudelitest eksisteerib ka kompaktsemaid versioone, aga nende võimekus on märgatavalt madalam. Täpsemalt on uuritud antud töö raames V3 ja R1 mudeli erinevusi ning treenimisvõimalusi peatükis 5.1.2.

3.2.2 Sünteetilised andmed

Lisaks Kruusmaa koostatud RaRa andmestikule kasutame valimi suurendamiseks sünteetilisi andmeid. Inspiratsiooniks on selle juures 2024. aastal avaldatud artikkel (Guan jt.), kus vigade reprodutseerimiseks kasutati tõenäosuslikku lähenemist. Selle tarvis kõrvutati kaks sama sisuga teksti - üks ilma vigadeta ning teine vigadega. Seejärel leiti iga sümboli jaoks tõenäosused, et see asendatakse OCR protsessi käigus mingi teise sümboliga [27].

Tõenäosuste leidmiseks peame kõigepealt joondama tekstid, mille jaoks kasutame Bio.Align raamistiku PairwiseAlignment klassi [28]. Joondamisloogika põhineb erinevatel bioinformaatikas kasutatavatel algoritmidel, näiteks nagu Needleman-Wunsch algoritm. Needleman-Wunsch algoritmi kasutatakse selleks, et joondada valgu- või nukleotiidijär-

jestusi ning leida mutatsioone kahe jada vahel (lisandused, eemaldused, vahetused) [29]. PairwiseAligner abil joondame inimparandatud- ja OCR teksti ning lisame sümboli "-" kohtadesse, kus sõnad on eripikkustega. Joondamist illustreerib järgnev näide.

Inimparandatud tekst	**OCR tekst**	**Joondatud inimparandatud tekst**	**Joondatud OCR tekst**
lõpulikult kindlustatud ja alaline Rahwa	lõpulikult ja alaline Rahwa	lõpulikult kindlustatud ja alaline Rahwa	lõpulikult - - - - - - - - - - - - -ja alaline Rahwa

Joonis 1. *Näide joondatud inimkontrollitud- ja OCR tekstist*

Seejärel loeme algoritmi abil kokku tähevastavused inimparandatud tekstis ehk mitu korda on korrektsetes tekstides olevatele tähtedele vastanud mingi täht vastavatest OCR tekstidest. Asendustõenäosused salvestame JSON faili, et neid kasutada korrektsetesse tekstidesse vigade lisamiseks.

Inimparandatud tekst	**OCR tekst**	**Asendustõenäosused tähele "a"**
Wiljandimaa wabataht-likud	Wiljandimao wabataht-!ikud	a: 0.83 o: 0.17

Joonis 2. *Näide asendustõenäosustest tähele "a"*

Järgmiseks võtame veebikoorimise meetodil (*web scraping*) Vikitekstid[2] veebilehekülgelt erinevate 19. ja 20. sajandi romaanide tekste ning lisame juurde RaRa andmestiku inimparandatud tekstid. Seejärel kasutame salvestatud asendustõenäosusi, et genereerida nendesse tekstidesse sünteetilisi vigu (Lisa 5).

Lisaks genereeritakse sünteetilisi vigu veel juhuslikkuse alusel. Esmalt valib algoritm ühe kolmest käigust: lisandus, asendus või kustutamine. Seejärel valitakse suvaline täht eesti keele tähestikust ja asenduse või lisamise korral lisatakse teksti juhuslikkuse alusel valitud kohale ning kustutamise korral eemaldatakse sama loogika järgi. Vigade sissetoomise ülemine piir on toodud järgmise valemiga, kuid 0.1 ühe asemel on kohati olnud kasutusel ka natukene suuremaid ja väiksemaid väärtusi.

$$\text{vigade ülempiir} = max(1, int(len(tekst) \cdot 0.1)) \tag{3.1}$$

[2]Vikitekstid `https://et.wikisource.org/wiki/Esileht`

Samuti kasutame enimlevinud OCR vigade asendamist, kus oleme käsitsi tekstide kontrollimise käigus avastanud mõned levinumad vead, näiteks "ii"-st saab "ü", ning asendame need täpselt samamoodi juhuslikkuse alusel tekstidesse. Need metoodikad on aga vähem esindatud ning sünteetiliste vigade andmestikus on kokku vaid umbes 400 näidet, mis on sellisel moel saadud.

3.2.3 Valideerimine

OCR-tekstide paranduste kvaliteedi hindamiseks kasutatakse peamiselt kahte näitajat: CER (vigaste tähtmärkide protsent tekstis) ja WER (vigaste sõnade protsent tekstis) ning seetõttu on need ka meie töös enimkasutatavad analüüsitööriistad [30]. Kuna mõlema mõõdiku puhul on tegu protsendiga, siis saame nende abil kõige parema ülevaate sellest, kui täpne mõni konkreetne mudel andmestiku peal on. CER ja WER saadakse vastavalt vigaste tähtede või sõnade arvu jagamisel kogu tekstis olevate tähtede või sõnade arvuga. Vigade leidmiseks kasutame Levenshteini kaugust, mille väärtus näitab kui mitu lisandust, asendust või eemaldust on vaja ühes tekstis teha, et see viia identsele kujule teisega [31].

$$CER = \frac{V + E + L}{T} \cdot 100\% \tag{3.2}$$

$$WER = \frac{V + E + L}{S} \cdot 100\% \tag{3.3}$$

kus:

- V = Vahetuste arv tekstis (valed tähed/sõnad)
- E = Eemalduste arv tekstis (puuduvad tähed/sõnad)
- L = Lisanduste arv tekstis (üleliigsed tähed/sõnad)
- T = Originaaltekstis sisalduvate tähtede arv
- S = Originaaltekstis sisalduvate sõnade arv

Esmalt arvutame CER-i OCR teksti ja inimparandatud teksti vahel. Peale mudeli treenimist ja testimist arvutame need väärtused uuesti, aga seekord mudeli pakutud paranduse ning inimparandatud teksti vahel. Seejärel lahutame need esimesest arvutusest saadud CER-ist ning saame CER muutuse (valem 3.4). Selle abil on lihtne näha millised parandused lähevad halvemaks (negatiivne muutus) ning millised paremaks (positiivne muutus). Sama loogika kehtib WER väärtuste puhul.

$$CER_{muutus} = CER_{vana}(GT, OCR) - CER_{uus}(GT, parandus) \quad (3.4)$$

Samuti arvutame CER ja WER paranemist üle kogu andmestiku, selleks lahutatakse kogu andmestiku keskmisest CER väärtusest OCR teksti ja inimparandatud teksti vahel andmestiku keskmine CER väärtus mudeli paranduse ja inimparandatud teksti vahel. Analoogselt tehakse sama WER väärtustega.

$$CER_{paranemine} = CER_{keskmine}(GT, OCR) - CER_{keskmine}(GT, parandus) \quad (3.5)$$

Kuigi ei ole kindlat mõõdupuud, mille järgi saaks öelda kust jooksevad hea ja halva CER-i piirid, siis trükitud teksti korral loetakse tavaliselt kõrge kvaliteediga tekste selliseks, mille CER jääb alla 2%, keskmise kvaliteediga tekstideks on CER vahemikus 2-10% ning üle 10% CER-iga tekstide kvaliteeti loetakse kesiseks [30] (Lisa 6).

Objektiivse hindamise kõrval on olulisel kohal ka tunnetuslik hinnang. Tihti ütlevad esimesed 10-15 näidet suures pildis ära, kas mudel on paranduste kvaliteedi poolest konkurentsivõimeline või mitte. Siiski, alati ei saa enne põhjalikumat analüüsi mudeli kohta suuremaid järeldusi teha, sest osad keelemudeli parandused on küll inimsilmale asjalikud, aga CER ja WER mõõdikute poolest kesised ja vastupidi.

Lisaks sellele, et meie mudel on vabavaraline ning sobib kasutamiseks ka piiratud õigusega tekstide parandamiseks, võtsime eesmärgiks jõuda paranduste kvaliteedis vähemalt samale tasemele kõige värskemate kommertsmudelitega: ChatGPT ja DeepSeek, milleks on töö kirjutamise hetkel ChatGPT-4o ja DeepSeek V3, R1. Selle jaoks kasutame ChatGPT ja DeepSeek APIt, millel laseme parandada oma testandmestikku. Kuna nende keelemudelite puhul tuleb ette üsna palju olukordi, kus lisaks parandatud tekstile tagastatakse teksti ees veel lause või sõna, näiteks "Siin on parandatud tekst", siis on need tulemuste ausaks võrdlemiseks eemaldatud. Samuti on standardiseeritud kõikide mudelite lõikes jutumärgid. Ajalehtedes sageli kasutatakse jutumärkidena „ ja ”, keelemudelid aga sagedamini kasutavad " sümboleid. Kuna teksti sisu poolest see midagi ei muuda, siis on CER ja WER tulemuste ühtlustamise ja võrdlemise eesmärgil jutumärgid standardiseeritud.

3.3 OCR tekstide ja paranduste kvaliteedi hindamine

OCR tekstide ja paranduste kvaliteedi hindamiseks katsetame kahe erineva lähenemisega. Esmalt treenime preemiamudelit, mis hindab vastuste kvaliteeti. Preemiamudeleid kasutatakse peamiselt RLHF treenimisel mudeli tagasisidestamiseks. Mudel tagastab tõenäosusliku hinnangu, kui lähedal on vastus ideaalsele. Treenimisel antakse sisendiks viip ning eelistatud ja tagasilükatud väljund ja treenitakse SKM-i. Mudel peaks selle järgi õppima, milliseid vastuseid hinnata kõrgemalt, milliseid mitte [32]. Kasutame treenimisel sama viipa, mida paranduste mudeli treenimiseks. Eelistatud väljundiks on inimparandatud tekst või heal tasemel keelemudeli parandus. Tagasilükatud väljundiks on algne OCR tekst või halb parandus, kus suhteline Levenshteini kaugus inimparandatud tekstist on üle 0.03. Andmestikus kasutusel olevad parandused on meie poolt treenitud paranduste mudeli väljundid.

Teise lahendusena kasutame peenhäälestamist. Selle käigus treenime kaks mudelit kahe erineva viibaga, kuna peenhäälestatud mudelid hakkavad täitma kahte erinevat ülesannet.

Esimene mudel hindab paranduste kvaliteeti. Selleks antakse võrdlemiseks mudelile OCR tekst ning keelemudeli tehtud parandus. Paranduse hindamise väljund on tõenäosus, mis väljendab seda, kui tõenäoliselt on parandus parem OCR tekstist. Seda on oluline väljendada just tõenäosusena, kuna mõni parandus on algsest OCR-ist selgelt halvem või parem, mõni aga võrdne või väga vähesel määral erinev. Paranduste puhul ei ole oluline ainult vigade arv tekstis (CER), vaid ka see, et parandused lähtuksid algtekstist. Kui näiteks lasta CER-i ennustada paranduse puhul, mis on keeleliselt korrektne, kuid algsest OCR tekstist üldse ei lähtu, oleks CER hinnang positiivne ning jätaks mulje, et paranduse käigus teksti kvaliteet paranes. Seetõttu on oluline anda mudelile sisendiks kaks teksti.

Teine mudel proovib ennustada ette antud OCR teksti CER-i, samuti laseme samal mudelil hinnata ka paranduste CER-i, kuid eeldame, et see ei ole edukas ning ühtlasi ei treeni me mudelit konkreetselt parandusi hindama. RaRa saab kasutada mudelit selleks, et tuvastada tekstide ligikaudset kvaliteeti ning otsustada, kas neid lasta paranduste mudelil parandada või on teksti kvaliteet nii kehv, et seda on parem parandada käsitsi. Lisaks on antud mudel oluline RaRa jaoks veel väljapool OCR tekstide järeltöötlust, kuna mudeli abil on võimalik luua ülevaade kogu andmestiku CER tasemest, leida kriitilised tekstid, mis vajavad uuesti OCR tehnoloogia abil töötlemist ning tekstid, mida uuesti töötlema ei pea.

Viibad mõlemale peenhäälestatud hindamismudelile leiab lisast 3.

3.3.1 Uus hindamisskaala

Preemiamudeli puhul kujuneb hindamisskaala treenimise käigus, mis tuleneb sellest, et preemiamudel peaks õppima andma välja tõenäosust vahemikus 0..1, kas tekst on korrektne või mitte ning seetõttu meie skaalat ei vali ega mõjuta. Peenhäälestatud mudeli hindamisskaala on vahemikus 0-100. CER-i ennustamise puhul on treenimisel hindamisskaala väärtusteks võetud CER väärtused (OCR teksti ja inimparandatud teksti vahel) ning need ümardatud täisarvuks, kuna komakohtadega täpsus ei ole konkreetse ülesande lahendamiseks vajalik.

Tõenäosuse, et parandus on parem kui OCR tekst, ennustamise puhul on hindamisskaala arvutatud CER vähenemise protsendi põhjal (valem 3.6). Kui paranduse CER vähenemine on 100 ehk näide parandati õigeks, on hindamise väljund 100. Kui CER vähenemine on -100 või enam, siis on väljund 0. Kui on vähenemine 0, on väljund 50. Ülejäänud väärtused on skaleeritud vastavalt.

$$CER_{\text{vähenemine}} = \frac{CER_{muutus}}{CER_{vana}(GT, OCR)} \cdot 100\% \tag{3.6}$$

3.3.2 Mudeli valik

Kuna nii paranduste hindamisel kui ka CER-i ennustamisel on samuti oluline keeleline kontekst, kasutame ka hindamismudelite jaoks Llammas mudelit.

3.3.3 Valideerimine

Kuna meile teadaolevalt ei ole sarnase hindamisskaala ja lähenemisega keegi hinnanud OCR tekste ja parandusi, siis kasutame oma tulemuste valideerimiseks erinevaid metoodikaid. Esmalt koostame testandmestiku analoogselt treeningandmestikule, kuhu arvutame tegeliku CER-i ja tõenäosuse, et parandus on parem kui algne OCR tekst. Hindamisel saab selle abil võrrelda, kui täpne on treenitud mudel ning arvutada keskmise erinevuse mudeli ennustuste ning tegelike väärtuste vahel.

Kuna hindamise puhul on oluline, et mudel klassifitseeriks teksti õigesti ehk saaks aru kui OCR-i või paranduse kvaliteet on väga hea, väga halb või keskmine, siis lisaks täpsele arvulisele võrdlemisele hindame ka, kas mudel kategoriseerib tekste õigesti ehk kas ennustatud hinne on õiges vahemikus. See annab mudelile pisut eksimisruumi, kuid meie

ülesande kontekstis vigu juurde ei tekita. Näiteks paranduste kvaliteedi hindamisel on eelkõige oluline, et mudel eelistaks parandusi ainult siis, kui need päriselt teksti kvaliteeti parandavad. Selle puhul aga ei ole oluline kas tõenäosuslik väljund oli 60 või 90, vaid kas tekst on parem või mitte (binaarne klassifikatsioon). Samamoodi CER-i ennustamisel on pigem oluline suurusjärk, mis määrab ära teksti kvaliteedi. Selleks määrame CER vahemike järgi tasemed ning vaatame kui täpne on mudel õige taseme määramisel.

Võrdlusena analoogselt paranduste mudelile, laseme sama ülesannet täita ka ChatGPT-4o mudelil ning DeepSeek V3 mudelil. Eeldame, et kuna need mudelid ei ole meie andmestiku peal treenitud, siis on nende sooritus ilmselt kesine, kuid kasutame seda siiski võrdlusena, et näha üldist keelemudelite võimekust konkreetse ülesande lahendamiseks.

3.4 Tehnoloogiavalikud

3.4.1 Programmeerimiskeel

Programmeerimiskeelena kasutame läbivalt Pythonit, sest see on üks enimkasutatavaid keeli masinõppes ja andmetöötluses. Python on intuitiivne ja lihtsa süntaksiga ning pakub palju erinevaid teeke ja raamistikke, mis kiirendavad ja lihtsustavad arendusprotsessi masinõppes. Ka Hugging Face veebiplatvorm, kus on ühtlasi avaldatud valitud mudel Llammas, on Pythoni põhine. Samuti on Pythonil väga hea kogukond ja internetitugi ning probleemide või küsimuste korral leiab suure tõenäosusega internetist vastuse [33].

3.4.2 Keskkonnad

Mudelite treenimiseks ja testimiseks kasutame kahte Tallinna Tehnikaülikooli hallatavat arvutuskeskkonda.

Esimene neist on AI-Lab, kus teeme esimesed katsetused peenhäälestamisega. Kasutame selleks AI-Labi kõige võimsamaid sõlmi - ai-lab-01 ja ai-lab-04, millel on 24 GB videomälu (VRAM) ja protsessoritel vastavalt 24 tuuma / 48 lõime ning 32 tuuma / 64 lõime [34].

Suuremate ja ressursinõudlikumate ülesannete jaoks, nagu treenimine, vajame aga AI-Labist võimsamat keskkonda, et vältida takerdumist mälupuudusest tulenevate probleemide taha. Treenimise ajal hoiab VRAM endas kõiki mudeli parameetreid, mille arv ulatub tavaliselt miljonitesse, ning 24GB mälust jääb selliste ülesannete puhul väheks. Seetõttu kasutame ka võimsat, kuid tasulist Teadusarvutuste Keskust. Teadusarvutuste Keskus pakub sõlmi, millel on minimaalselt 64 tuuma ja 128 lõime, suurematel 128 tuuma ja 256

lõime, ning see on meie tegevusteks piisav.

Mõlemad keskkonnad on Slurm põhised. Slurm (*Simple Linux Utility for Resource Management*) on arvutiklastrite haldussüsteem, mida kasutatakse tööde jagamiseks mööda arvutiressursse laiali, tööde monitoorimiseks ning tööde järjekorra haldamiseks [35].

Kõikideks tegevusteks, mis ei hõlma endas mudeli treenimist või testimist, kasutame Google Colab pilvepõhist Python'i arenduskeskkonda [36]. See on mõeldud erinevate masinõppe ning andmetöötlusega seotud tegevuste jaoks ning kuna oleme seda varem erinevate õppeainete raames kasutanud, siis on valik selle kasuks loogiline. Enamiku tegevuste jaoks kasutame Colabi T4 GPU-d, mis katab kõik meie vajadused. Kuna Colab on Google'i kasutajaga seotud, siis saame omavahel koodi ja tulemusi jagada ning vajadusel üksteise märkmikke täiendada.

3.4.3 Analüüsimeetodid

Igale treenimistsüklile järgneb tulemuste analüüs, mille maht oleneb enamasti sellest, kui häid tulemusi mudel kindlates testides saab. Paranduste mudeli puhul arvutame esmalt alati summaarse-, keskmise- ja mediaan CER-i ja WER-i, et hinnata mudeli täpsust testandmestikul. Sageli aitab see otsustada, kas konkreetse mudeliga tasub edasi töötada. Hindamismudelite puhul arvutame kõigepealt keskmise- ja mediaanerinevuse tegelike ja ennustatud hinnete vahel.

Treening- ja testandmestikud on meil vastavalt JSON- ja CSV-formaadis. Kuna suurem osa andmetöötlus hõlmab endas CER-i ja WER-i arvutamist, siis enim kasutame JiWER teeke [37]. JiWER teegiga leiame CER ja WER väärtused kasutades sisseehitatud funktsioone ning korrutame tulemused 100-ga, et saada protsendid.

Andmete visualiseerimiseks ning CSV-failidega töötamiseks kasutame pandas, matplotlib ja seaborn teeke [38, 39, 40].

4. Andmestik

Peatükis 4.1 tutvustame Kruusmaa poolt koostatud andmestikku üldisemalt. Peatükis 4.2 räägime lähemalt sellest, kuidas Kruusmaa andmestikku enda töös kasutame. Samuti tuleb selles peatükis juttu sünteetilistest vigadest (4.2.1). Treenimiseks kasutatud andmestikud on kaardistatud peatükis (4.2.2). Räägime ka Kruusmaa koostatud andmestikus esinenud probleemidest (4.2.3), testandmestiku parandamisest (4.2.4) ning viimaseks võrdleme enda andmestikku teistes sarnastes töödes kasutatud andmestikega peatükis 4.2.5.

4.1 Ülevaade RaRa andmestikust

Krister Kruusmaa poolt koostatud ja filtreeritud andmestikus on 7410 teksti 179 erinevast perioodikaväljaandest [4]. Keskmine sõnade arv teksti kohta on umbes 99 sõna, kõige lühem tekst on üks sõna ja pikim 3235 sõna pikk. Kõige varasem tekst on aastast 1821 ning kõige hilisem aastast 2014. Enamik tekste jäävad 20. sajandi esimesse poolde, kus suurim osakaal on tekstidel aastatest 1918 ja 1919, kuna neid konkreetseid tekste parandati massdigiteerimise projekti raames [2].

Andmestiku keskmine CER on 11.77% ning keskmine WER 36.64%. Suurem osa tekste on CER väärtusega alla 20% ja WER väärtusega alla 50%. Kolm näidet on andmestikus juba ideaalse CER ja WER väärtusega ehk inimparandatud tekstiga samad. Halvim CER väärtus on 88.44% ja halvim WER väärtus 200%. Halvima WER näite puhul on OCR tekstis kaks korda rohkem sõnu kui inimparandatud tekstis, halvima CER näite puhul on tekst vale, kuna inimparandatud tekstis on kõik tähed suured tähed. Samas ei saa selle lõputöö ja andmestiku kontekstis seda näidet õigeks lugeda, kuigi sisulise tähenduse poolest on see õige.

4.2 Töös kasutatav andmestik

Algsest andmestikust on eemaldatud kirillitsa tähti sisaldavad tekstid ning saksakeelsed ajalehed. Samuti on eemaldatud liiga pikad tekstid, näiteks algse andmestiku pikim tekst, mis on üle 3000 sõna, ei ole Llammas keelemudelile jõukohane, kuna maksimaalne *tokenite* arv on 4096. Töös kasutusel olevasse RaRa andmestikku jäi alles 7146 teksti, kõige pikem tekst on 508 sõna, keskmine sõnade arv umbes 90. CER ja WER väärtuste puhul märkimisväärseid muudatusi ei esine. Juhuslikkuse alusel on jaotatud andmestik test- ja treeningandmeteks. Treeningandmestikus on 5145 näidet ja testandmestikus 2001

näidet. Nende mõlema CER ja WER väärtused püsisid samuti samas suurusjärgus, mis algselt üle kogu andmestiku. Sõnade arv treeningandmestiku OCR tekstides kokku on 468 541 ning inimparandatud tekstides 463 062. Testandmestikus on vastavalt 181 679 ja 179 939 sõna.

Andmestikku on lisaks kaasatud 555 näidet 1991. aasta Eesti Päevalehest, mille transkribeeris üks RaRa poolsetest juhendajatest (Lisa 7). Nende Eesti Päevalehe tekstide hulgast on eemaldatud ühe terve lehekülje jagu näiteid, sest lehekülg sisaldas endas ainult rootsikeelset telekava. Lisaks on andmestikku lisatud ligikaudu tuhat näidet välja võetud pikkade tekstide hulgast, mis on manuaalselt tükeldatud ja inimparandatud tekstidega kokku sobitatud.

4.2.1 Sünteetiliste vigade andmestik

Suurem osa sünteetiliste andmete jaoks kogutud inimparandatud tekste pärineb eelnevalt mainitud vabade alliktekstide internetikogumist Vikitekstid. Seal saavad samuti soovijad OCR teksti parandada, kuid suur eelis DIGAR-i ees on see, et tekstid läbivad mitu kontrollringi ning sellest tulenevalt on neil küljes staatusesildid (Õigsus tõendamata, Õigsus tõendatud, Heakskiidetud, Probleemne) [41]. Meie andmestikus on kasutusel ainult sellised tekstid, mille õigsus on tõendatud või heakskiidetud. Kokku on veebikoorimise teel saadud tekstides lauseteks tükeldatuna 26 934 näidet, millele lisanduvad RaRa töötaja transkribeeritud 555 näidet. Lisaks genereeritakse vigu RaRa treeningandmestiku 5145 inimparandatud tekstidesse, et paremini säilitada andmestiku eripärasid.

Vikitekstide andmestikul ja RaRa andmestikul on kaks peamist erinevust. Esiteks on sünteetiliste vigadega andmestikus inimparandatud tekstid täielikult kontrollitud ja korrektsed. Teiseks on eelnimetatud andmestikus esindatud näited lausekaupa, samas kui RaRa andmestikus need ei ole tingimata lausepõhised, vaid enamjaolt esindatud ajalehelõikudena ehk koosnevad mitmest lausest.

Vikitekstidel põhineva sünteetilise andmestiku keskmine CER on 8.5% ja keskmine WER 43.1%. Suurim CER on 100% ja WER 200%. Ideaalse CER ja WER väärtusega ehk 0% on 686 näidet. Keskmine sõnade arv umbes 15.5 sõna näite kohta. Pikim tekst on 203 sõna ja lühim üks sõna pikk. Kuna andmestikus on näited lühemate tekstidega, siis näidete arv on suurem, kuid sõnade arv on samas suurusjärgus RaRa andmestikuga ehk OCR tekstides kokku 418 057 sõna ning inimparandatud tekstides 422 742 sõna.

4.2.2 Treeningandmestikud

Töö käigus on kombineeritud RaRa andmestikust saadud treeningandmestikku, RaRa inimparandatud tekstidest (GT ehk *ground truth*) saadud sünteetiliste vigade andmestikku ja sünteetiliste vigadega Vikitekstide andmestikku erinevatel viisidel, kuna mudeleid on treenitud erinevat suurusjärku ja sünteetiliste andmete osakaaluga andmetega. Märgitud on lisaks näidete arvule sõnade arv, kuna näidete pikkused varieeruvad suurel hulgal ning sõnade arv annab parema ülevaate sellest, kui palju andmestik tegelikkuses suurenes.

Tabel 1. *Treeningandmestikud*

Metoodika	**Andmestik**	**Näiteid**	**Sõnu OCR**	**Sõnu GT**
Paranduste peenhäälestus väheste andmetega (Llammas FT-5k)	RaRa treeningandmestik	5145	468 541	463 062
Paranduste peenhäälestus kombineeritud andmetega (Llammas FT-13k)	RaRa treeningandmed algkujul ja sünteetiliste vigadega, Vikitekstid (3795 näidet)	13 715	976 661	976 032
Paranduste peenhäälestus rohkete andmetega (Llammas FT-43k)	RaRa treeningandmed algkujul ja sünteetiliste vigadega kahekordselt, Vikitekstid	43 389	1 820 331	1 828 122
Paranduste eelistuste suunamine ehk DPO ainult RaRa	RaRa treeningandmed koos Llammase või meie tehtud mudeli ennustustega	5145	468 541	463 062
Paranduste eelistuste suunamine ehk DPO kombineeritud andmetega	RaRa treeningandmed, filtreeritud sünteetilised andmed koos Llammase või meie tehtud mudeli ennustustega	11 319	942 613	941 817

Jätkub...

Table 1 – *Jätkub...*

Tabel 1. *Treeningandmestikud*

Metoodika	**Andmestik**	**Näiteid**	**Sõnu OCR**	**Sõnu GT**
Kvaliteedi hindamine preemiamudeliga	RaRa treeningandmed, sünteetilised filtreeritud kujul andmed koos meie mudeli tehtud ennustustega	11 576	998 767	980 948
Kvaliteedi hindamise peenhäälestus: CER ennustamine	RaRa treeningandmed, sünteetilised andmed	9231	529 221	-
Kvaliteedi hindamise peenhäälestus: paranduste hindamine	RaRa treeningandmed parandustega, sünteetilised andmed	10 300	937 145	-

4.2.3 Probleemid andmestikuga

Nagu sünteetiliste vigade alampeatükis märgitud, siis on RaRa andmestikus enamjaolt tekstid, mis koosnevad rohkem kui ühest lausest. Vaadeldes töid, kus tekstiparanduste kvaliteet on läbivalt hea, paistab silma, et tihti on tekstid maksimaalselt 2-3 lauset pikad [42]. Oma andmestiku tekste ei lühenda me põhjusel, et see oleks liiga aeganõudev. Kuna kohati OCR tekstid ei meenuta väljanägemise poolest isegi inimkeelset teksti ning koosnevad lihtsalt erinevatest kirjavahemärkidest ja tekstisümbolitest, siis oleks inimparandatud ja vigaste tekstide lausehaaval kokkusobitamine konkreetse andmestiku puhul keeruline. Teine võimalus oleks tekste kokkusobitada käsitsi, aga see oleks lihtsalt väga aeganõudev.

RaRa andmestiku puhul on suur probleem see, et iga inimparandatud tekst ei ole tegelikult täies mahus või korrektselt parandatud. Töö käigus avastasime, et paljud inimparandatud tekstid algavad küll korrektselt, aga poole pealt on parandamine pooleli jäetud ning tekst on vigane. Seetõttu on töö üheks eesmärgiks ka testandmestiku kõik inimparandatud tekstid õigeks parandada, et paremini hinnata mudelite tegelikku sooritust.

Kuna RaRa andmestikus on tekste pea 200 aasta ulatuses, siis on keelekasutus kohati väga erinev. See teeb ülesannet raskemaks selle poolest, et alati ei saa öelda, et näiteks ’v’

tuleks asendada 'w'-ga. Kuna keelekasutus on kahe viimase sajandi jooksul olnud pidevas muutumises, siis võivad tekstid üksteisest suuresti lausestuse ja grammatika poolest erineda, olenevalt väljaande aastast. Samuti on erinevates Eesti maakondades erinev keelekasutus ning sama sõna võib esineda andmestiku lõikes mitmes erinevas vormis, mis parandamise puhul on eriti keeruline, kuna iga näite puhul tuleb kontekstist lähtuvalt valida õige vorm. Siin ei aita ka ajalehe nime viibas ette andmine, kuna DIGAR-is esinevas andmestikus on erinevaid väljaandeid palju rohkem võrreldes meie treeningandmetega, seetõttu peaks mudel ise olema suuteline vajalikku konteksti tuvastama.

Lisaks sellele, et ajalooliste tekstide kirjapilt ja kvaliteet on üle andmestiku varieeruvad, on ka OCR tekstide kvaliteet vägagi kõikuv. Osade tuvastatud tekstide kvaliteet on väga hea ning parandusi on vaja teha vähe, samal ajal kui mõni teine tekst on ka inimsilmale täiesti loetamatu ja seetõttu ka praktiliselt parandamatu. Sellest tulenevalt võib mudeli sooritus olenevalt näitest olla väga erinev. Seetõttu jaotame testandmestiku vigade kontsentratsiooni ja tekstide pikkuste põhjal erinevateks raskusastmeteks ning tulemustes hindame lisaks üldisele sooritusele ka edukust vastavalt algsele vigade kontsentratsioonile või teksti pikkusele.

OCR mudelitele tekitab kohati ajalehtede puhul probleeme tulpade tuvastamine, kui tulbad on üksteisele liiga lähestikku lehele paigutatud. Näiteks esineb andmestikus tekste, mis ei ole lugedes üldse loogilised, kuna OCR on kokku pannud kahe tulba tekstid läbisegi. See omakorda raskendab keelemudeli jaoks parandamist ning konteksti mõistmist (Lisa 8).

4.2.4 Testandmestiku parandamine

Testandmestiku parandamise tulemusena muutus keskmine CER testandmestikul -0.77% ning keskmine WER -2.33%. Vigaseid tekste oli CER-i poolest 1034 (51.67%), millest suurim muutus oli -32.74%, WER-i poolest 890 (44.48%), millest suurim muutus -70.98%. Põhjus, miks CER-i poolest on rohkem vigaseid tekste on see, et CER arvestab ka osaliselt õigete sõnadega, sest osad tähed on inimparandatud tekstile vastavad, aga WER loeb ka kasvõi ühe tähevea korral valeks terve sõna.

Parandatud tekstidest muutus keskmiselt parandust vajanud tekstide CER -1.49% ja WER -5.24%. Kõige enim vajasid parandamist pikad teksti (rohkem kui 110 sõna), kus 75.80% tekste vajas parandamist, vähim väga lühikesed tekstid (vähem kui 15 sõna), kus parandamist vajas 15.25% tekste. See ilmselt tuleneb sellest, et pikkade tekstide puhul on suurem tõenäosus, et parandaja jätab mõned sõnad või laused parandamata.

Lisaks selgus, et siiski on andmestikus mõned võõrkeelsed (saksa, prantsuse, soome,

rootsi) tekstid, mis on ilmunud eestikeelsetes perioodikaväljaannetes. Neid ei eemaldatud, kuna need on olnud osa eestikeelsetest väljaannetest ja seetõttu loeme need eestikeelsete ajalooliste tekstide osaks. Samuti ei parandanud me võõrkeelseid tekste, kuna me ei ole selleks piisavalt pädevad. Endiselt esineb mõningates inimparandatud tekstides veel sõnu algsel OCR kujul, kuna nende välja lugemine väljaande skaneeringult ei olnud võimalik.

Kuna test- ja treeningandmestik on jaotatud juhuslikult kaheks, võib eeldada, et vigu tähemärkides (CER) ja sõnades (WER) on tegelikkuses treeningandmetes umbes sama palju, kui kontrollitud testandmetes. Samuti esineb samal eeldusel treeningandmetes sama palju vigaseid inimparandatud tekste. See on ka meie töö üks puudus, kuna ei ole võimalik välja selgitada kui palju vigased inimparandused treenimist mõjutavad, sest selle jaoks oleks vaja ära kontrollida ka kogu treeningandmestik, mille jaoks ei ole käesoleva lõputöö raames piisavalt ressursse. Samuti tuleks puudusena mainida asjaolu, et sünteetilised andmed on genereeritud RaRa andmestiku vigade põhjal, mistõttu mõjutavad vigased inimparandused sealgi vigade kontsentratsiooni ning tõenäosusi.

4.2.5 Andmestiku võrdlus teiste sarnaste tööde andmestikega

Tulemuste objektiivseks võrdluseks on oluline mainida ära sarnaste tööde andmestike suurusjärgud ning kvaliteedid. Turu ülikooli soomekeelses OCR andmestikus on keskmine CER 9% ja WER 28% ning tekstid ajavahemikus 1836-1918 [12]. Andmestikus on sõnu OCR tekstides 449 088 ja inimparandatud tekstis 461 305 Inglisekeelse andmestikuga töö puhul, kus on samuti peenhäälestatud Llama-2 keelemudelit ning tulemused on väga positiivsed, on enne parandusi keskmine CER 7.52% ja WER 24.23%. Tekstid on aastate vahemikus 1800-1900. Treeningandmestikus on 238 471 sõna OCR tekstides ja 233 430 inimparandatud tekstis. Võrdlusena on selles andmestikus 10 400 näidet, mis on kaks korda suurem, kui meie treeningandmestik, kuid sõnade arvult ligi kaks korda väiksem kui meie andmestik [42].

Meie andmestik on teiste sarnaste tööde andmestikest halvema kvaliteediga nii CER kui ka WER parameetrite poolest. Samuti on meie andmestikus esindatud suurem ajavahemik. Oluline erinevus on näidete pikkuse osas, mis võib mõjutada paranduste kvaliteeti ning seetõttu on tulemustes eraldi jaotus ka pikkuste põhjal. Lisaks ei esine teistes töödes üldjuhul poolikuid lauseid, mida meie andmestikus on palju. See kõik omakorda raskendab ülesannet, kuna peale selle, et eesti keele puhul on tegu on väikekeelega ning meie töö on meile teadaolevalt esimene taoline suurem katsetus eesti keeles, on meie andmestiku üldine kvaliteet märgatavalt kehvem.

5. Paranduste mudel

Selles osas tutvustame lähemalt OCR vigade parandamiseks mõeldud mudelite treenimise ja testimise protsesse. Esmalt räägime lähemalt töövoost (5.1), mis hõlmab endas sobiva viiba leidmist (5.1.1), olukorra kaardistamist (5.1.2), üldist tööprotsessi 5.1.3 ning viimaks kirjeldame täpsemalt eelistuste suunamise meetodit peatükis 5.1.4. Peatükis 5.2 toome välja ja analüüsime erinevate mudelite tulemusi. Kõik tulemused on toodud peatükis 5.2.1, mida analüüsime peatükis 5.2.2. Viimaks selgitame ja põhjendame testandmestiku raskusastmeteks jaotamist peatükis 5.2.4, mida analüüsime peatükis 5.2.5.

5.1 Töövoog

Paranduste mudeli treenimise töövoog on nii peenhäälestamise kui ka eelistuste suunamise metoodika puhul samasugune. Enne mudelite treenimist katsetame treenimata Llammasega, et hinnata mudeli üldist võimekust, leida sobivaim viip meie ülesande ja treenimise kontekstis ning tuvastada suuremad potentsiaalsed probleemid ning veakohad.

Peenhäälestamisel on kasutusel kolm erinevat mudelit: Llammas-base, Llammas ja Llammas-GEC. Viimane neist on treenitud spetsiaalselt kirjavigade parandamiseks. Nii Llammas-base kui ka Llammas-GEC mudelite peal teostame ainult peenhäälestamist, kuna eelistuste suunamine on eelkõige mõeldud vestlusrobotite treenimiseks.

Keelemudeli parameetrite osas teeme väiksel hulgal katsetusi. Selleks kasutame 50 halvimate parandustega testandmestiku näidet, mille puhul testime ja võrdleme tulemusi erinevate parameetritega. Põhjus, miks me täisandmestikul parameetritega ei katseta on see, et see oleks väga ajakulukas ja ressursimahukas tegevus, kuna parameetrite muutmisel on võimalusi ja kombinatsioone palju ning raske on ette ennustada, kuidas mingi muudatus tulemusi mõjuda võib. See tähendaks seda, et tuleks parameetreid muuta vähehaaval ja ühekaupa ja seetõttu peaks seda tegema kümneid kordi. Seetõttu eelistame erinevate parameetrite testimiseks väiksemat alamhulka andmestikust ning katsetame vaid väga väheste parameetrite muutmist.

Meie treenitud mudelite kvaliteedi hindamiseks võrdleme pidevalt tulemusi treenimata Llammase, ChatGPT-4o, DeepSeek V3 ja DeepSeek R1 mudeli tulemustega. See aitab kaardistada, mis võiks üldiselt olla keelemudelite võimekus konkreetse ülesande lahendamiseks ning kas treenimise tagajärjel on parandused kvaliteetsemad või mitte.

5.1.1 Sobiva viiba leidmine

Esmase katsetusena jooksutame kogu RaRa andmestiku (nii treening- kui ka testandmestiku) treenimata Llammase mudeliga, et kaardistada ära üldine baasvõimekus. Genereerimisel kasutame iga näite puhul viite erinevat viipa. Vastavad viibad on koostatud eesmärgiga, et iga viiba puhul lisandub mingi oluline detail antud näite või ülesande kohta. Kõige tavalisem viip on ilma kontekstita ja annab ainult juhise teksti parandada, järgmisel lisandub teksti väljaandmise aasta, seejärel täiendavad juhised, näiteks ära lisa sõnu juurde.

Kõige sobivama viiba leidmiseks hindame käsitsi väikest hulka tekste ning arvutame CER väärtused, mille tulemusena selgub parim viip, mis on kasutusel kõikides treenimistes. Selgub, et detailne viip ei anna mudelile eelist, just vastupidi, ning lühike ja konkreetne juhend aitab kaasa kvaliteetsematele parandustele. Samuti, kuna treenime mudelit täitma väga spetsiifilist ülesannet, ei ole vaja reegleid ja juhiseid väga detailselt ette kirjutada, sest mudel peaks ülesande iseloomust ja reeglitest ise treenimise käigus arusaama omandama.

Katsete tulemusena selgub, et ChatGPT-4o mudeli puhul mängib viip palju suuremat rolli, kui Llammase puhul. Testimine on tehtud kahe viibaga - üks sama, mis on kasutusel treenimisel, aga seda on täiendatud lausega, et mudel tagastaks ainult parandatud teksti. Seda seetõttu, et antud keelemudelil on komme lisaks relevantsele väljundile tagastada veel juurde selgitavat teksti. Kuna aga antud ülesande kontekstis me seda ei soovi ning iga sõna väljundis maksab, siis on vaja täiendavate juhistega seda välistada. Llammase mudeli puhul sellist probleemi ei esine. Teine viip on sisukamate juhistega ning suunab keelemudelit ülesannet konkreetsel viisil lahendama. Kasutatud viibad on leitavad lisast 3.

Kuna pikem ja põhjalikum viip annab kommertsmudelite puhul märgatavalt paremad tulemused, siis DeepSeek V3 ja R1 mudelite testimisel on kasutatud ainult pikemat viipa. Lisaks on API-de puhul probleemne, et vastuste andmises võib esineda tõrkeid. Näiteks on mõned väljundid täiesti tühjad ning sellistes olukordades laseme keelemudelil sama näidet uuesti parandada. Samuti on üksikuid olukordi, kus mudel ei anna parandust väljundiks ja tagastab selle asemel hoopis teksti, et parandamine ei olnud võimalik, sest näide on parandamatu või vajab täiendavat konteksti. Sellised näited on säilitatud tulemustes, kuna kajastavad keelemudelite tegelikku võimekust ja probleemsust.

5.1.2 Olukorrast ülevaate saamine

Meie poolt peenhäälestamata Llammase tulemustest selgub, et mudeli peamine probleem on sõnade väljamõtlemine ja lisamine. Eriti sageli juhtub see olukordades, kus näide on

poolik lause, näiteks algab poole sõna pealt või lõppeb kooloni või komaga. Sellisel puhul lisab Llammas peaaegu alati pooliku sõna asemele mõne eestikeelse ja antud konteksti sobiva sõna. Koma ja kooloni puhul üldiselt jätkab Llammas lauset, tihti isegi lisab otsa mitu lauset. Lisandused on küll enamjaolt loogilised ja konteksti sobivad, kuid antud ülesande raames valed. Probleeme on mudelil ka Eesti kohanimede ja isikute äratundmisega. See on arusaadav, kuna vajaks selle jaoks üsna laialdast treenimist ajalooliste isikute ning kohanimedega. Samuti ei saa eeldada, et mudel suudab tuvastada ja õigeks parandada tavainimeste nimesid.

Veel on oluliseks Llammas mudeli puuduseks sõnade kaasajastamine. Mudel on grammatikas tugev ning valdab eesti keelt hästi, mistõttu sageli kaasajastab vanema keelekasutusega sõnu, kuna tänapäevases keelekasutuses need ongi valed. Sama probleem esineb kohanimedega - näiteks on andmestikus Narva kujul Narowa, mis keelemudelile võib tunduda kui OCR viga, mitte ajalooline eripära.

Võrreldes kommertsmudelitega on Llammase eelis eesti keele tundmine, sest lisaks meie treenimisele on Llammas eelnevalt treenitud ligikaudu 5 miljardit eestikeelset *tokenit* sisaldava andmestikuga [21]. Lisaks, kuna Llammase viibas täiendavaid juhiseid ei ole, siis proovib mudel alati parandusi teha, samal ajal kui ChatGPT ning DeepSeek jätavad rohkem sõnu või tekste tervikuna algsele OCR kujule. See ei ole otseselt halb omadus, kuna CER ja WER näitajate puhul on eelistatud pigem algne OCR vorm kui täiesti välja mõeldud sõna või tekst. Samas peab aga parandamise ja mitteparandamise vahel olema teatav tasakaal, kuna mudel, mis suurema osa ajast tagastab OCR tekste algsel kujul, on samuti kasutu.

DeepSeek mudelite V3 ja R1 võrdlusest selgub, et R1 mudeli testimine on väga palju kallim ja võtab kordades rohkem aega. Seetõttu esmalt testime terve andmestiku V3 mudeliga, seejärel testime esimesed 308 (15.4%) näidet R1 mudeliga. Nende näidete võrdlemisel selgub, et R1 mudeli tulemused on halvemad ning edasist testimist me eelmainitud kulukuse tõttu ei tee.

DeepSeek mudeleid on võimalik lokaalselt peenhäälestada. Selle juures on aga antud töö kontekstis probleeme. Esmalt R1 mudeli, mille jaoks on vastavalt olemas väiksemaid ja kompaktsemaid mudeleid, on andmestiku koostamine keeruline ülesanne. Andmestikku tuleks lisada arutluskäik, mis paranduste ülesande puhul arutleks läbi kõik vead, mida mudel parandab, või jätab alles ja seejuures igat oma otsust põhjendab. See on väga ajamahukas, kuna vajaks detailset analüüsimist, kontrollimist ja käsitsi parandamist ning seetõttu antud lõputöösse ei mahu. Teisena oleks võimalik peenhäälestada DeepSeek teisi mudeleid, mis sellist andmestikku ei vaja. V3 mudelid on kõik väga suuremahulised

ning nende treenimiseks meil arvutusvõimekust ei ole. Üle jääb veel DeepSeek-LLM-7b[1], mis suuruse poolest on võrreldav Llammasega, kuid kuna see anti välja aastal 2023 ning on treenitud inglise- ja hiina keelega ja seetõttu ei näe me sellel mingit eelist Llammase treenimisega võrreldes. Samuti konkreetse mudeli üksikute näidetega testimisel joonistub välja selgelt kehvem paranduste kvaliteet võrreldes V3 mudeliga.

2025. aasta kevadel on ilmunud ka Google Gemma 3 mudelite seeria, mis toetab 140 erinevat keelt, sealhulgas eesti keelt. Kuna aga ilmumisaeg jääb antud töö lõpufaasi, siis ajapuudusest tingituna ei ole selle mudeli potentsiaali OCR tekstide järeltöötluse ja tekstide kvaliteedi hindamise puhul testitud [43].

5.1.3 Üldine tööprotsess treenimisel

Treenimise protsess algab treening- ja testkoodi koostamisega. Selle jaoks uurime sarnaste tööde peenhäälestuskoode ning Llammas mudeli enda peenhäälestuskoode. Katsetamise tulemusel, kus toimub pidev treenimine, testimine ning tulemuste analüüs, on eesmärk jõuda meie jaoks kõige paremini sobiva peenhäälestamise koodini. Lisaks CER-i ja WER-i analüüsimisele hindame tulemusi ka subjektiivselt, et märgata võimalikke probleeme või leida, milliseid näiteid oleks vaja rohkem treeningandmetesse tuua.

Parameetrid, mis on kasutusel testimisel: temperatuur=0.7, top_p=0.1, top_k=40. Mudelite treenimisel on 5 epohhi. Sisendil on maksimaalne *tokenite* arv 4096, mis on ühtlasi ka Llama-2 mudelitel suurim lubatud arv. Väljundi puhul lubatud uute *tokenite* arv on dünaamiline. See tuleneb pigem harva esinevast olukorrast, kus mudel jääb mingi näite puhul ühte sümbolisse, sõnapaari või üldse poolikusse sõnasse kinni ning kordab seda. Kinni jäämise tulemusena genereerib mudel täis lubatud *tokenite* arvu. Selleks, et sellised üksikud näited liigselt keskmist ei mõjutaks, on lubatud genereerida sisendi ehk OCR teksti *tokenite* arvust maksimaalselt 10 *tokeni* võrra pikem väljund. Kuna mudel peaks ainult vigu parandama ja sõnu ning lauseid mitte juurde lisama, siis antud piir on meie hinnangul mõistlik. Andmestikus esineb küll näiteid, kus inimparanduses on mõni sõna rohkem kui OCR tekstis, kuid sellistel juhtudel ei ole OCR suutnud puudu olevat sõna tuvastada ning ei saa eeldada, et mudel oskaks või peaks sinna midagi ise õigesti asemele pakkuma.

Kokku on peenhäälestatud kolme erineva treeningandmestikuga, mille sisu on kajastatud peatükis 4.2.2. Andmestikud annavad parimaid tulemusi juhul, kui näited on treeningandmestikus suvalises järjekorras ehk sünteetilised vead ja RaRa originaaltekstid esinevad

[1] Hugging Face DeepSeek https://huggingface.co/deepseek-ai/deepseek-llm-7b-base

segamini. Samuti on osad mudelid treenitud nii, kus alguses koosneb andmestik ainult RaRa näidetest ning seejärel on jätkatud treenimist sünteetiliste andmetega. Sellise lähenemise miinuseks on, et keelemudel kipub esmast treenimist unustama ning tugevamalt tulevad esile parandamismustrid viimasest treenimisest. Seetõttu sellised mudelid tulemustes ei kajastu.

5.1.4 Eelistuste suunamine

Eelistuste suunamisel on treenimisel kasutusel nii treenimata Llammas kui ka meie poolt juba eelnevalt peenhäälestatud Llammas mudelid. Kuigi peamiselt kasutatakse eelistuste suunamist eelkõige vestlusrobotite keelekasutuse ning tekstide kokkuvõtmise oskuste parandamiseks, siis on tegelikult treenimise struktuur ka meie ülesande jaoks sobilik. DPO (*Direct Preference Optimization*) treenimisel on sisendiks viip, mis koosneb küsimusest või ülesandest ning kahest vastusest: üks eelistatud vastus ning üks tagasilükatud vastus. Meie kasutame ülesande edasiandmiseks täpselt sama viipa, mida ka peenhäälestamisel ehk OCR teksti ning käsklust seda parandada. Eelistatud vastuseks on meie puhul inimparandus või mõningate näidete korral ka hea parandus. Selliste näidete korral on tagasilükatud vastuseks inimparandatud tekst, kuhu on sisse toodud sünteetilisi vigu. Tagasilükatud vastused saame parasjagu treenimisele mineva mudeli enda ennustustest treeningandmestikule. Osadest treeningandmestikest on välja filtreeritud sellised näited, kus eelistatud ja tagasilükatud vastused on liiga sarnased ehk kus parandus on liiga hea kvaliteediga.

Puudus meie DPO treenimisel on peenhäälestatud Llammase treeningandmestik, kuna mudel on samade näidetega juba treenitud, mistõttu on tehtud parandused paremad kui oleksid täiesti uute andmetega. Kuna aga täiesti uue andmestiku koostamine oleks väga ajamahukas, siis kasutame siiski RaRa ja sünteetiliste andmete treeningandmestikku. Treenitud on nii sünteetiliste andmetega kombineeritult kui ka ilma, nagu on kirjeldatud peatükis 4.2.2.

Eelistuste suunamisega treenimisel on tulemused oodatust halvemad. Kõikide katsete peale kokku õnnestus reaalsuses arvestataval tasemel ainult üks, ning isegi selle mudeli paranduste kvaliteet jääb märgatavalt alla peenhäälestatud mudelitele. Õnnestunud tähendab siinkohal seda, et mudel reaalselt üritab tekste parandada ning parandused vastavad üldjoontes ikkagi originaaltekstile. Treeningandmestiku suurendamisel sünteetiliste andmetega jääb aga mudel mingisse esimese näite sõnapaari kinni ning kordab seda iga näite puhul. Viimase DPO katsetusena on treenitud Llammast tulemuste parandamise eesmärgil sellise RaRa andmestikuga, millele oleme rakendanud eelnevalt mainitud filtreerimise loogikat, kuid ka see ebaõnnestus.

Eelistuste suunamise ebaõnnestumises kahtlustame kahte erinevat põhjust. Kuna eelistuste suunamine on peamiselt viis, kuidas viia mudelite vastuseid rohkem vastavusse sellega, mida kasutaja ootab, siis kahtlustame, et see on natukene kõrvalekalduv lähenemine sellest, mida meie palume oma mudelitel teha. Kuna DPO on mõeldud pigem avatud küsimuste lahendamiseks, nagu näiteks tekstide kokkuvõtmine, dialoogi pidamine jne, siis võib-olla on ühe õige vastusega ülesanne, nagu tekstis vigade parandamine, DPO skoobist väljas. Kuna mudelile pole teada põhjused, miks üks vastus on eelistatud ja teine mitte, siis on võimalik, et treenimisel haarab meie mudel kinni mingitest juhuslikest nüanssidest, mida peab eelistatud ja mitte-eelistatud tekstide vahel määravateks faktoriteks. Loomingulisemate ülesannete puhul ei pruugi mõningal valehinnangul olla erilist olulisust, kuid kuna tekstides vigade parandamine nõuab eelkõige täpsust, siis ei saa meie sellist varieeruvust ja juhuslikkust endale lubada.

Teine oluline faktor ebaõnnestumisel võib olla see, et eelistuste suunamine ei olegi sobilik meetod väiksemate keelemudelite treenimiseks. 2022. aastal avaldatud artiklis leidsid Bai jt., et eelistuste suunamisest saavad peamiselt kasu just suuremad keelemudelid [44]. Kuna meie kasutame 7 miljardi parameetriga mudelit, siis on reaalne võimalus, et meie mudel on lihtsalt eelistuste suunamisest kasu saamiseks liiga väike.

5.2 Tulemused

Peenhäälestamisel selgub, et Llammas-base on sama andmestikuga treenimisel Llammase vestlusroboti peenhäälestamise tulemustest alati halvem. Seetõttu ei ole järgnevates tulemustes kajastatud Llammas-base tulemusi.

Llammas-GEC tulemused peenhäälestamise käigus ei paranenud võrreldes algsete Llammas-GEC mudeli tulemustega, mistõttu ei ole samuti need tulemused järgnevates tabelites kajastatud.

Parameetritega katsetustest selgub, et enim positiivset mõju kehvadele näidetele avaldab top_p=0.9. Kordustega ehk kinni jäänud näidete puhul repetition_penalty=1.15 märkimisväärselt tulemusi ei paranda. Samuti selgub, et kui samu näiteid pidevalt uuesti genereerida, võib juhtuda, et samade parameetritega annab mudel erinevaid väljundeid. See tuleneb ilmselt kõrgest top_p või temperatuuri (meie puhul läbivalt 0.7) väärtusest, mis suurendavad vastuste varieeruvust. Uuesti genereerimise võimalust ning parameetrite vahetamist paremate tulemuste saamiseks kasutame mudelite koostööl, mis on täpsemalt selgitatud peatükis 7.

Kuna eelistuste suunamise puhul saab enamik katseid lugeda ebaõnnestunuks, siis

järgnevates tulemustes kajastub kõige paremini õnnestunud eelistuste suunamisega treenitud mudel, milleks on ainult RaRa andmestikuga treenitud Llammas.

5.2.1 Tulemused kogu andmestiku peal

Kõik tulemused on saadud RaRa testandmestikuga testimisel. Parameetrid, mis kajastuvad tabelites, on CER ja WER muutuse põhjal. Seda seetõttu, et lihtsalt paranduse CER ja WER tase väga head ettekujutust sellest ei anna, kas paranduste käigus andmete kvaliteet paranes või mitte. Negatiivne muutus tähendab, et kvaliteet läks halvemaks ning positiivne, et läks paremaks. Järgnevalt on selgitatud tulemuste statistikas kasutusel olevad väärtused.

- Keskmine – kogu andmestiku ulatuses keskmine CER/WER muutuse väärtus
- Mediaan – kogu andmestiku ulatuses mediaan CER/WER muutuse väärtus
- Parim muutus – kõige suurem positiivne muutus kogu andmestiku ulatuses
- Halvim muutus – kõige suurem negatiivne muutus kogu andmestiku ulatuses
- Muutus >= 0 – näidete osakaal protsendina, kus CER/WER muutus >= 0
- Ideaalne – inimparandustega identsete näidete osakaal protsendina

Tabelites on paksus kirjas märgitud iga parameetri võrdluses parim tulemus, et paremini eristada, milline mudel on üldiselt kõikide parameetrite lõikes parim.

CER tulemused

Tabel 2. *CER muutuste statistika*

Mudel	**Keskmine**	**Mediaan**	**Parim muutus**	**Halvim muutus**	**Muutus >= 0**	**Ideaalne**
ChatGPT-4o	-2.51%	-1.16%	21.74%	-278.85%	38.03%	0.30%
DeepSeek V3	-4.13%	-1.46%	21.74%	-492.33%	33.78%	0.35%
Llammas	-8.90%	-4.27%	16.98%	-143.67%	23.74%	0.00%
Llammas-GEC	-14.33%	-1.55%	12.20%	-82.42%	33.08%	0.00%
Llammas FT-5k	-1.86%	0.00%	**50.00%**	**-80.24%**	52.67%	**1.75%**
Llammas FT-13k	**-1.37%**	**0.17%**	32.08%	-128.26%	**55.07%**	1.20%
Llammas FT-43k	-1.95%	0.13%	31.11%	-195.92%	53.82%	1.55%

Jätkub...

Jätkub...

Mudel	Keskmine	Mediaan	Parim muutus	Halvim muutus	Muutus >= 0	Ideaalne
Llammas DPO	-9.72%	-3.12%	18.87%	-99.23%	26.99%	0.00%

WER tulemused

Tabel 3. *WER muutuste statistika*

Mudel	Keskmine	Mediaan	Parim muutus	Halvim muutus	Muutus >= 0	Ideaalne
ChatGPT-4o	2.01%	0.98%	100.00%	-328.57%	63.37%	0.30%
DeepSeek V3	1.66%	1.95%	100.00%	-428.46%	66.47%	0.45%
Llammas	-7.29%	-2.33%	100.00%	-209.52%	42.93%	0.00%
Llammas-GEC	-11.83%	-3.13%	37.5%	-188.88%	41.63%	0.00%
Llammas FT-5k	9.30%	7.69%	100.00%	-294.41%	**89.16%**	**1.75%**
Llammas FT-13k	**10.46%**	8.77%	100.00%	**-140.00%**	87.41%	1.20%
Llammas FT-43k	9.65%	**9.09%**	100.00%	-289.69%	85.56%	1.55%
Llammas DPO	-8.77%	-1.69%	100.00%	-326.48%	44.68%	0.00%

5.2.2 Paranduste mudeli tulemuste analüüs

Parima mudeli välja selgitamine arvuliste parameetrite põhjal lihtne ei ole, kuna meie poolt peenhäälestatud Llammas FT-5k ja Llammas FT-13k on üsnagi võrdsed. Sünteetiliste andmete lisamine tõstis mudeli keskmisi- ja mediaanväärtusi, kuid ainult RaRa andmestikuga treenitud mudel on parem näidete täiesti korrektseteks parandamisel ning samuti parima CER muutuse poolest.

Võrdluses kommertsmudelitega ning algse Llammase tulemustega, on kõik meie peenhäälestatud mudelid paremad. ChatGPT-4o ja DeepSeek V3 puhul on tulemused ühtlasemad, kuna viibas on suunatud mudelit lauseid ja sõnu samaks jätma, kui parandamine võimalik ei ole, mis tähendab, et CER ja WER muutused ei ole nii drastilised. Väga negatiivsete muutustega näiteid on pigem üksikuid. Suur erinevus on näidete osakaalus, mille kvaliteet läks paremaks või säilis.

Võrreldes Llammase algsete tulemustega, on treenimise käigus tulemused märgatavalt paranenud. Erandiks on Llammas DPO, mille puhul tulemused läksid halvemaks. Seetõttu hindame, et eelistuste suunamine ei ole antud ülesande lahendamiseks sobiv metoodika. Lisaks tuleb puudusena märkida, et paremate tulemuste saavutamiseks võiks mudel olla treenitud kvaliteetsema eelistuste suunamise andmestikuga ning treenimise kombineerimisega näiteks preemiamudeliga või kasutades inimtagasisidet.

Llammas-GEC puhul on tulemused samuti kehvemad Llammase vestlusroboti tulemustest. GEC mudeliga on katsetused tehtud eelkõike seetõttu, et esmapilgul võivad tunduda OCR vigade ja kirjavigade parandamine üsna sarnaste ülesannetena. Tulemustest aga selgub, et reaalsuses on olukord vastupidine. Kirjavigade puhul lähtutakse üldjuhul juhuslikest klaviatuuril kirjutamisest tekkinud tähevigadest või levinumatest kirjavigadest. OCR tekstide puhul aga selliseid vigu esineb pigem harva ning vead on üldiselt seotud tähtede ja sõnade valesti tuvastamisega. Samuti mängib rolli ajalooliste tekstide puhul kulumine ja keerukad fondid, mistõttu vahepeal on konteksti põhjal vaja ise sõnu kokku panna ja puhtalt tähtede asendamisest ei piisa.

Tulemuste puhul on selgelt märgata, et WER tulemused on palju positiivsemad, kui CER tulemused. See tuleneb ilmselt kahest erinevast faktorist. Esiteks on algne WER müratase andmestikus palju suurem, kui CER puhul, mistõttu on WER parameetrite mõttes võimalik teha rohkem parandusi. Samuti saab olla näiteks algne WER -200%, kui OCR tehnoloogial on üks sõna tükeldatud mitmeks erinevaks ning seetõttu pealtnäha väikse paranduse sisse toomisel võib WER drastiliselt paraneda. CER saab antud andmestikus samuti olla üle 100%, kuid algsel kujul selliseid tekste ei esine. Seetõttu on WER muutused palju suurema väärtuse ning kaaluga kui CER muutused. CER tulemuste puhul on paranemiste protsendid üldjuhul pigem väikesed ning samal ajal halvenemiste protsendid võivad kergelt muutuda väga suureks, mistõttu kalduvad tulemused keskmise poolest negatiivseks.

Teine põhjendus kõrgemateks WER väärtusteks on see, et teataval määral on sõnavigu tähevigadest lihtsam parandada. Näiteks on üheks selliseks olukorraks näited, kus tekstis on suurem osa sõnu vigased, kuid vead on väga lihtsad. CER muutuse mõttes on paranemine sellises olukorras väike, WER muutuse mõttes aga suur.

Peenhäälestatud mudelite puhul on CER ja WER erinevused väga suured. Kommertsmudelite puhul on tulemused ühtlasemad ning muutused võrreldes peenhäälestatud mudelitega umbes kaks korda väiksemad. Ühtlased tulemused CER ja WER poolest pole aga konkreetse ülesande puhul eriti hea näitaja, sest nagu selgub tulemuste tabelist, siis ühtlane tähendab siin ühtlaselt kehv. Peenhäälestatud mudelite korral aga vigaste sõnade osakaal väheneb treenimise tulemusel. Seda ilmselt seetõttu, et isegi kui mudel ei suuda

tervet teksti õigeks parandada või mingite paranduste puhul teeb vigu, siis üldine WER muutus võib kokkuvõttes olla ikkagi positiivne, kuna näiteks lihtsamad või enamlevinumad vead sõnades parandatakse ära.

5.2.3 Mudelite võrdlus ja eripärad

Järgnevalt toome konkreetseid paranduste näited parimate mudelite lõikes. Töös on illustreeritud lühikesed näited, kuna pikkades tekstides vigade ja paranduste hoomamine on keeruline ning oluline on anda edasi mudelite paranduste kvaliteeti ja eripärasid ning selleks ei ole pikki näiteid ilmtingimata vaja.

Üks näide, mis illustreerib hästi paranduste eripärasid, on näide, mis on ühtlasi kogu andmestiku ja mudelite võrdluses parima CER muutusega (joonis 3). Tegu on tekstiga, kus algne müratase on väga kõrge. Parima paranduse on teinud arvuliste väärtuste põhjal Llammas FT-5k, halvima Llammas FT-43k. Samas, kõik meie poolt treenitud mudelid on suutnud tuvastada sõna "Eesti" ning õigesti kasutanud läbivalt "w"-d, mitte "v"-d. FT-5k on kõige paremini säilitanud algse OCR teksti olemuse, FT-13k on proovinud rohkem iga sõna eraldi parandada, FT-43k on sõnu juurde lisanud. ChatGPT on parandanud neid sõnu, mida pidas arusaadavaks, kuid kõik parandused on valed. DeepSeek tagastas teksti algsel kujul.

OCR tekst	C%#tt nwlxmifli ttwlüfmtfe ajii* liite forK
Inimparandus	Eesti wabariigi walitsemise ajutine kord.
FT-5k	Eesti wabariigi walitsuse ajutise asu
FT-13k	Eesti rahwa tulewiku küsimuse asjus liitriikide kokku
FT-43k	Eesti rahwusliku wäerinna kaitse wäerinna korraldamise asjus
ChatGPT-4o	C%#tt maailma tuleviku ajalugu liite forK
DeepSeek V3	C%#tt nwlxmifli ttwlüfmtfe ajii* liite forK

Joonis 3. *Mudelite võrdlus parima CER muutusega näite puhul*

Teise näite puhul on ChatGPT teinud ideaalse paranduse (joonis 4), samal ajal kui peenhäälestatud mudelid on teinud ühe olulise vea ning Viljandi esimeseks täheks asendanud "W". Seda ilmselt seetõttu, et "w" tähte kasutati aktiivselt kuni 1920ndate aastate keskpaigani ning suur osa meie treeningandmestiku näiteid pärineb vahetult sellele eelnevatest aastatest. DeepSeek V3 mudeli parandus on konkreetse näite puhul halvim, kuna pole Viljandit kui kohanime ära tundnud. Sõna "gümnaasiumi" parandamise puhul ühelgi mudelil probleeme ei tekkinud, mis näitab ühtlasi seda, kui oluline on konteksti tunnetus antud ülesande juures, kuna tänu eelnevale sõnale, kus on viidatud haridusele, ning järgnevale sõnale, kus on õpilased mainitud, saab teha järelduse, mis vahepealne vigane sõna võiks tegelikult olla.

OCR tekst	Vitjondi Haridusseltsi giiwwwogumi õpilasele
Inimparandus	Viljandi Haridusseltsi gümnaasiumi õpilasele
FT-5k	Wiljandi Haridusseltsi gümnaasiumi õpilasele
FT-13k	Wiljandi Haridusseltsi gümnaasiumi õpilasele
FT-43k	Wiljandi Haridusseltsi gümnaasiumi õpilasele
ChatGPT-4o	Viljandi Haridusseltsi gümnaasiumi õpilasele
DeepSeek V3	Võtma Haridusseltsi gümnaasiumi õpilasele

Joonis 4. *Mudelite võrdlus ideaalse ChatGPT parandusega näite puhul*

Kolmanda näitena on tekst, kus on pisivead - ekslikud suured tähed ning üks sõna, mis vajab sisulisemat parandust (joonis 5). DeepSeek on teinud selle puhul ideaalse paranduse, samuti FT-5k. ChatGPT-l on jäänud üks suur täht parandamata ning FT-43k-l kaks suurt tähte. FT-13k on küll suured tähed õigeks parandanud, aga sõna "lahkunud" on saanud vale vaste.

OCR tekst	Lehtedest käis teade läbi', st Tartu ülikookiist mitmed Professorid licchkunud. Nüüd Kir
Inimparandus	Lehtedest käis teade läbi, et Tartu ülikoolist mitmed professorid lahkunud. Nüüd kir
FT-5k	Lehtedest käis teade läbi, et Tartu ülikoolist mitmed professorid lahkunud. Nüüd kir
FT-13k	Lehtedest käis teade läbi, et Tartu ülikoolist mitmed professorid surnud. Nüüd Kir
FT-43k	Lehtedest käis teade läbi, et Tartu ülikoolist mitmed Professorid lahkunud. Nüüd Kir
ChatGPT-4o	Lehtedest käis teade läbi, et Tartu ülikoolist mitmed professorid lahkunud. Nüüd Kir
DeepSeek V3	Lehtedest käis teade läbi, et Tartu ülikoolist mitmed professorid lahkunud. Nüüd kir

Joonis 5. *Mudelite võrdlus ideaalse DeepSeek parandusega näite puhul*

Neljanda näite puhul on tegemist ülimalt keeruka OCR tekstiga, mille õigeks parandamine ilma alustekstita on pea võimatu (tabel 4). Parima tulemuse on saavutanud Llammas FT-13k, kus mitmed sõnad tekstis on kas osaliselt või tervenisti õigeks parandatud. Samuti ei ole mudel lisanud sõnu kohtadesse, kus parandamine on olnud liigselt keerukas ning seetõttu on säilinud OCR kuju. Napilt positiivse CER muutusega on ka DeepSeek. Teised parandused ei ole olnud edukad. ChatGPT ei ole parandust tagastanud, kuna kvaliteet on liiga halb. Nagu näha, siis kõik Llammase mudelid on teinud põhimõtteliselt suvalisest tähtede ja sümbolite jadast enam-vähem loetava teksti. Samas on nii vigase OCR teksti puhul raske inimparandusega identset teksti automaatsete paranduste teel saada ning tihti on just see põhjuseks, miks CER ja WER väärtus on algse tekstiga võrreldes palju kehvemad, kuigi visuaalselt tekst justkui paraneb.

Tabel 4. *Mudelite võrdlus keerulise OCR näite puhul*

OCR tekst	fttlt, peaaegu ttaau ’n iugifu.gu;e rasu!olematirfc Umbega. Iseapõt>i’Ujt ieltefß nüiffe fee artmtb olctoai, et cftnircfj-c Fõne nt tite Fiillflli p arterit fet ofmtb. tPieie ci tabafs fcCDeft riitte jÖpeTbaba, ci Stuguft 9krilt imitgifPgujt .miitingu fõ-net ootafib. ftebt ifir
Inimparandus	kult, peaaegu nagu mingisuguse rahulolematuse tundega. Peapõhjust selleks näikse see annud olewat, et esimehe kõne mitte küllalt parteilik ei olnud. Meie ei tahaks sellest mitte järeldada, et sotsialdemokraadid August Reilt mingisugust miitingu kõnet ootasid. Leht kir
FT-5k	Kui, peaaegu taas nagu suurgu’gu’e raskeolematu umbega. Iseapost, et kõneleja oli, ei tekkinud selle arwust midagi. Kõne oli ometi täiesti kõlbmatu. Mitte sellepärast ei tahaks sellest rääkida, et kõneleja oli, waid sellepärast, et kõne oli ometi täiesti kõlbmatu. Mitte sellepärast ei tahaks sellest rääkida, et kõneleja o
FT-13k	siis, peaaegu taas nagu mingisugune raskusolematuse Umbega. Iseapistust, milleks nüüd see artikb olctoai, et cftnircfjc Fõne nt tite Fiillflli parterit set ootas. Meie ei tahaks sellest mitte jääda, ei luguks 9krilt imitgifPgujt .miitingu fõnet ootafib. Peet
FT-43k	Kuulda, peaaegu taand-lik-uguse raskeolematusega. Iseäranis huwitaw oli selle artikli lõpp, mis arwustaja Fõne nii täielikult parandust selles osas leidis olewat. Meie ei tahaks sellest mitte järeldada, et kõneleja kõne mitte kõige paremat parandust selles osas leidis olewat. Meie ei tahaks sellest mitte järeldada, et kõneleja kõne mitte kõige paremat parand
ChatGPT-4o	Kahjuks ei suuda ma sellisel kujul olevat teksti parandada, kuna OCR tekst on väga moonutatud ja mulle ei ole teada piisavalt konteksti, et see õigesti taastada.
DeepSeek V3	fttlt, peaaegu täau’n iugifu. gu;e rasu!olematu Umbega. Iseäpäit’Ujt ieltefß nüidse fee artmtb olctoai, et cftnircfj-c Fõne nt tite Fiillflli p arterit fet ofmtb. tPieie ci tabafs fcCDeft riitte jÖpeTbaba, ci Stuguft 9krilt imitgifPgujt .miitingu fõ-net ootafib. ftebt ifir

Llammas FT-5k puhul halvima muutusega näide on olnud selline, kus OCR tekst on alanud "sõiduriist ilma hobuseta ...", paranduses on mudel sinna kinni jäänud ning pannud punkte

kuni *tokenite* arv täitus. Ülejäänud OCR teksti ei parandatud. Algse teksti kvaliteet on tegelikult väga hea - CER 2.39% ja WER 5.59%. Teised mudelid on parandanud selle näite puhul kas natuke paremaks või halvemaks, aga üldjoontes rahuldavalt. Llammas FT-13k puhul on samuti halvim näide selline, kus mudel on jäänud sõnapaari kordama. Teiste mudelite puhul konkreetse näitega seda probleemi ei esine ning parandused on adekvaatsed. Llammas FT-43k halvimaks paranduseks on inglisekeelne tekst.

Sellest tuleb välja üks peenhäälestatud mudelite halb eripära: mudelid jäävad kinni, hakkavad sümboleid või sõnu kordama ning sellest tingituna halvenevad CER ja WER väärtused väga drastiliselt. Siit tõuseb esile paranduste- ja hindamismudeli efektiivse koostöö olulisus, sest kui hindamismudel suudab automaatselt kõige kehvemad parandused üles leida, siis saab need lasta näiteks paranduste mudelil uuesti parandada, parameetreid muuta või hoopis algse OCR tekstiga asendada. Selle tulemusena andmestiku keskmine kvaliteet paraneb või jääb halvimal juhul samaks ning sellega koos paraneb ka kasutajakogemus.

CER ja WER muutuste erinevuste illustreerimiseks on näide, kus OCR kvaliteet on samuti pigem kehv, kus teksti algus on lihtsamini parandatav, kuid lõpp on segane (joonis 6). Llammas FT-5k ja FT-13k mudelite puhul on CER muutus negatiivne, kuid WER muutus positiivne. See tuleneb sellest, et OCR tekstis olev üleliigne "." on ära kaotatud, samuti on FT-13k õigesti parandanud sõnad "Seltsi" ja "kõrgemas", teised sõnad, aga on valed, mistõttu CER muutus on kokku tulnud negatiivne. ChatGPT puhul on CER muutus positiivne, sest vigade kontsentratsiooni tähemärkides on selgelt vähenenud, kuigi sõnad on ikkagi suures osas vigased. DeepSeek-i parandus on kehv, sõnu on juurde lisatud.

OCR tekst	Eesti . eiilunött Seitfi satt(n kvrgemaS WUusttKZsM-iH
Inimparandus	Eesti Helikunsti Seltsi Tartu kõrgemas muusikakoolis
FT-5k	Eesti sõjawägi Soomes suuremas wäljaõppe-kursusel
FT-13k	Eesti Põllumeeste Seltsi kõnekoosolek kõrgemas wäärtuses
FT-43k	Eesti . diplomaatias sai viimase kümne aasta jooksul kõrgemaks kohaks välisministeerium
ChatGPT-4o	Eesti Tööline Seitse aastat kõrgemas Kunstikoolis
DeepSeek V3	Eesti rahva nõukogu Seitsmes sessioon käesoleva aasta WUusttKZsM-iH

Joonis 6. *Mudelite võrdlus CER ja WER erinevuste näitamiseks*

5.2.4 Raskusastmed

Kuna üldistest tulemustest ei joonistu hästi välja mudelite eripärad ning kogu andmestiku peal saavutatud tulemuste põhjal parimat mudelit välja selgitada on raske, siis täpsemaks analüüsiks jaotame testandmestiku omakorda veel raskusastmeteks. Selle järgi on võimalik avastada keelemudelite piire ja paremini hinnata, milline peaks olema optimaalne

andmestik ning millised on reaalsed ootused keelemudelitele antud ülesande raames. Selle tulemusena on võimalik anda RaRale täpsem ülevaade meie keelemudelite võimekusest.

Raskusastmeteks jaotamiseks kasutame kolme erinevat parameetrit: testandmestiku OCR tekstide CER väärtused, WER väärtused ning tekstide pikkused sõnade arvu järgi. Eeldame, et teatud CER/WER väärtustest alates on tekstide parandamine mudelile pea võimatu, kuna OCR müratase on sellisel juhul nii suur, et näide on arusaamatu ja loetamatu. Samuti eeldame, et pikemate näidete puhul ei ole mudel suuteline nii hästi enam parandama kui lühikeste puhul.

Raskusastmete leidmiseks lähtume testandmestiku jaotustest. Andmed jaotame nelja kategooriasse nii, et igasse raskusastmesse jääks enam-vähem võrdne arv näiteid. Võrdleme eelnevatest tulemustest parimaid mudeleid raskusastmete kaupa: ChatGPT-4o, DeepSeek V3, Llammas FT-5k, FT-13k ja FT-34k. Välja on jäetud Llammas, Llammas DPO ja Llammas-GEC, kuna nende tulemused on teiste mudelite tulemustest selgelt halvemad. Kaasates kõik meie peenhäälestatud mudelid, saame paremini analüüsida andmete koguse ning sünteetiliste andmete mõju kategooriate kaupa. Järgnevas tabelis on toodud raskusastmete jaotused.

Tabel 5. *Raskusastmed*

Kategooria	**Vahemik**	**Näidete arv**
CER kategooriad		
Väga madal müratase	0% <= CER < 5.5%	518
Madal müratase	5.5% <= CER <= 9.5%	465
Keskmine müratase	9.5% < CER < 16%	518
Kõrge müratase	16% <= CER	500
WER kategooriad		
Väga madal müratase	0% <= WER <= 20%	494
Madal müratase	20% < WER < 34%	495
Keskmine müratase	34% <= WER <= 54%	508
Kõrge müratase	54% < WER	504
Teksti pikkuse kategooriad		
Lühikene tekst	0 <= sõnade arv < 24	486
Keskmise pikkusega tekst	24 <= sõnade arv <= 60	506
Pikk tekst	60 < sõnade arv <= 125	497
Väga pikk tekst	125 < sõnade arv	512

Raskusastmete tulemustes kajastuvad parameetrid on samad, mis olid kasutusel koond-

tulemustes. Paksus kirjas on iga kategooria lõikes parim tulemuse antud parameetri puhul. Olukorras, kus enam kui kolmel tulemusel on sama parim tulemus, ei ole seda paksus kirjas rõhutatud, kuna ei anna teiste ees enam märkimisväärset eelist.

CER raskusastmete statistika

Tabel 6. *CER raskusastmete tulemused*

Mudel	**Keskmine**	**Mediaan**	**Parim muutus**	**Halvim muutus**	**Muutus >= 0**	**Ideaalne**
Väga madal müratase						
ChatGPT-4o	-2.56%	-1.46%	4.59%	-93.14%	23.94%	0.58%
DeepSeek V3	-2.07%	-1.30%	**4.62%**	**-72.38%**	25.10%	0.39%
Llammas FT-5k	**-0.58%**	**0.17%**	4.46%	-80.24%	**61.00%**	**4.25%**
Llammas FT-13k	-0.95%	0.09%	4.60%	-95.90%	55.21%	2.70%
Llammas FT-43k	-1.10%	0.00%	4.56%	-85.11%	53.09%	3.67%
Madal müratase						
ChatGPT-4o	-1.42%	-0.93%	8.16%	-44.40%	39.78%	0.22%
DeepSeek V3	-1.23%	-0.81%	7.55%	**-41.68%**	39.35%	0.43%
Llammas FT-5k	-0.93%	0.14%	**8.93%**	-70.19%	55.27%	**1.08%**
Llammas FT-13k	**-0.65%**	**0.44%**	8.47%	-68.59%	**58.71%**	0.65%
Llammas FT-43k	-1.18%	**0.44%**	8.16%	-87.82%	57.63%	0.86%
Keskmine müratase						
ChatGPT-4o	-1.74%	-0.28%	13.40%	-145.18%	48.84%	0.19%
DeepSeek V3	-1.98%	-1.32%	13.16%	-74.03%	40.93%	0.58%
Llammas FT-5k	-2.25%	-0.24%	**14.29%**	**-57.09%**	48.65%	0.97%
Llammas FT-13k	**-1.45%**	0.00%	**14.29%**	-128.26%	52.90%	0.77%
Llammas FT-43k	-1.77%	**0.33%**	**14.29%**	-68.89%	**54.05%**	**1.16%**
Kõrge müratase						
ChatGPT-4o	-4.29%	-1.91%	21.74%	-278.85%	39.80%	0.20%
DeepSeek V3	-11.21%	-4.05%	21.74%	-492.33%	30.20%	0.00%
Llammas FT-5k	-3.66%	-0.78%	**50.00%**	**-71.88%**	45.80%	**0.60%**
Llammas FT-13k	**-2.37%**	**0.37%**	32.08%	-124.43%	**53.80%**	**0.60%**
Llammas FT-43k	-3.74%	0.00%	31.11%	-195.92%	50.80%	0.40%

WER raskusastmete statistika

Tabel 7. *WER raskusastmete tulemused*

Mudel	**Keskmine**	**Mediaan**	**Parim muutus**	**Halvim muutus**	**Muutus >= 0**	**Ideaalne**
Väga madal müratase						
ChatGPT-4o	-8.78%	-7.78%	14.44%	-96.43%	24.09%	0.40%
DeepSeek V3	-5.81%	-4.29%	14.44%	-93.43%	32.59%	0.61%
Llammas FT-5k	**1.88%**	**2.70%**	**20.00%**	-294.41%	**83.60%**	**4.25%**
Llammas FT-13k	1.41%	2.22%	18.18%	**-87.43%**	72.47%	2.63%
Llammas FT-43k	1.22%	2.36%	18.18%	-193.90%	75.91%	3.44%
Madal müratase						
ChatGPT-4o	-0.19%	0.00%	27.27%	-38.89%	54.34%	0.20%
DeepSeek V3	2.10%	2.50%	33.33%	**-33.33%**	67.88%	0.40%
Llammas FT-5k	7.62%	8.16%	33.33%	-89.89%	92.53%	0.61%
Llammas FT-13k	**8.07%**	8.62%	33.33%	-70.41%	**92.93%**	0.40%
Llammas FT-43k	7.84%	**9.40%**	33.33%	-71.11%	89.70%	**0.81%**
Keskmine müratase						
ChatGPT-4o	4.83%	5.93%	50.00%	-200.00%	81.30%	0.39%
DeepSeek V3	4.74%	5.88%	50.00%	-428.46%	82.48%	0.39%
Llammas FT-5k	9.71%	11.11%	50.00%	-77.99%	89.96%	**1.38%**
Llammas FT-13k	**12.00%**	**13.33%**	50.00%	-140.00%	**92.91%**	1.18%
Llammas FT-43k	11.33%	12.95%	50.00%	**-72.47%**	89.17%	**1.38%**
Kõrge müratase						
ChatGPT-4o	11.92%	13.37%	100.00%	-328.57%	88.69%	0.20%
DeepSeek V3	5.45%	7.96%	100.00%	-209.09%	82.14%	0.40%
Llammas FT-5k	17.82%	16.67%	100.00%	**-50.00%**	90.48%	**0.79%**
Llammas FT-13k	**20.13%**	**20.56%**	100.00%	-86.25%	**91.07%**	0.60%
Llammas FT-43k	18.00%	18.94%	100.00%	-289.69%	87.30%	0.60%

Teksti pikkuste raskusastmete statistika

Tabelis olevad parameetrid on CER muutuste põhjal.

Tabel 8. *Teksti pikkuste raskusastmete tulemused*

Mudel	**Keskmine**	**Mediaan**	**Parim muutus**	**Halvim muutus**	**Muutus >= 0**	**Ideaalne**
Lühike tekst						

Jätkub...

Table 8 – *Jätkub...*

Mudel	**Keskmine**	**Mediaan**	**Parim muutus**	**Halvim muutus**	**Muutus >= 0**	**Ideaalne**
ChatGPT-4o	-3.18%	-1.08%	21.74%	-278.85%	47.74%	0.62%
DeepSeek V3	-6.65%	-1.94%	21.74%	-221.95%	41.36%	1.03%
Llammas FT-5k	-3.61%	-1.16%	**50.00%**	**-71.88%**	46.71%	**4.12%**
Llammas FT-13k	**-2.47%**	**0.00%**	32.08%	-128.26%	**52.67%**	2.67%
Llammas FT-43k	-3.51%	**0.00%**	31.11%	-195.92%	50.62%	3.29%
Keskmise pikkusega tekst						
ChatGPT-4o	-2.04%	-1.08%	12.74%	-108.02%	41.30%	0.20%
DeepSeek V3	-3.21%	-1.30%	10.16%	-84.59%	37.94%	0.40%
Llammas FT-5k	-0.90%	0.43%	15.60%	-76.26%	57.11%	1.58%
Llammas FT-13k	**0.02%**	**0.71%**	**24.55%**	**-68.59%**	**61.46%**	0.99%
Llammas FT-43k	-1.15%	0.57%	23.02%	-76.26%	58.70%	**1.98%**
Pikk tekst						
ChatGPT-4o	-1.83%	-0.88%	14.41%	-87.44%	36.82%	0.00%
DeepSeek V3	-2.22%	-1.25%	17.29%	**-54.19%**	33.00%	0.00%
Llammas FT-5k	-0.72%	0.30%	20.00%	-75.48%	59.36%	**1.01%**
Llammas FT-13k	**-0.37%**	0.45%	**20.85%**	-81.04%	**60.76%**	0.80%
Llammas FT-43k	-0.41%	**0.47%**	20.16%	-65.50%	59.76%	0.80%
Väga pikk tekst						
ChatGPT-4o	-3.02%	-1.42%	10.73%	-93.14%	26.76%	**0.39%**
DeepSeek V3	-4.52%	-1.53%	6.78%	-492.33%	23.24%	0.00%
Llammas FT-5k	**-2.27%**	**-0.09%**	16.26%	**-80.24%**	**47.46%**	**0.39%**
Llammas FT-13k	-2.66%	-0.19%	17.36%	-124.43%	45.51%	**0.39%**
Llammas FT-43k	-2.76%	-0.20%	**19.44%**	-87.82%	46.29%	0.20%

5.2.5 Analüüs raskusastmete kohta

CER raskusastmete puhul on mudelite lõikes selge, et mida suurem on müratase, seda kehvemad on tulemused. Parimad tulemused on väga madala ja madala müratasemete puhul ning seal on ka rohkem ideaalselt parandatud näiteid. Kõrge müratasemete puhul on kontrastselt palju suuremad halvimate muutuste väärtused. Seda soodustab ilmselt, et kõrge müratasemete puhul on raskem eristada ja ennustada konkreetseid sõnu või tähevigu, mis omakorda soodustab keelemudelite puhul sõnade välja mõtlemist. Samuti on kõrgema müratasemega suurem tõenäosus, et peenhäälestatud mudelid jäävad kinni, eriti olukorras, kus OCR ei ole loetav ning on sisuliselt lihtsalt sümbolite ja tähtede jada. Parim muutus on

suurim kõrge müratasemе puhul, madalaim väga madala puhul, mis on ka loogiline, sest väga madala müratasemega on algne CER väärtus vaid paar protsenti, siis on ka suurim võimalik paranemise protsent tagasihoidlik.

WER raskusastmete puhul on tulemused vastupidised CER tulemustega. Mida kõrgem on müratase, seda kõrgemad on keskmiste ja mediaanide muutused. Küll aga on suurem ideaalselt parandatud näidete osakaal väga madala müratasemе puhul. Kommertsmudelite puhul on üllatav, et väga madala müratasemе puhul on muutuste keskmine ja mediaan tugevas miinuses. Peenhäälestatud mudelite puhul nii drastilist muutust ei toimu.

Peenhäälestatud mudelite puhul treeningandmete suurendamine sünteetiliste andmete arvelt mängib suurimat rolli just siis, kui müratase läheb kõrgemaks. Nii CER kui WER puhul on Llammas FT-5k parim väga madala müratasemе puhul, kuid kõikides teistes kategooriates muutub eelkõige Llammas FT-13k selgelt paremaks. Sellest saab järeldada, et rohkem treeningandmeid tagab parema toimetuleku raskemate näidete puhul, kuid lihtsamate näidete korral paranduste kvaliteet kergelt halveneb, kuna mudel proovib parandada liiga palju. Kumba mudelit eelistada, oleneb seetõttu parandamisele minevate tekstide müratasemest.

Teksti pikkuste puhul on paranduste kvaliteedile halva mõjuga väga lühikesed ning väga pikad tekstid. Parimad tulemused saavutab mudel tekstide peal, mille pikkus on 24 ja 125 sõna vahel. Lühikeste tekstide puhul on ilmselt suurim probleem see, et valedel parandustel on lühikeste tekstide korral suurem kaal. Samas on vastukaaluks lühikeste tekstide puhul lihtsam teha ideaalseid parandusi, sest on hea võimalus, et mudelil on vaja teha vaid mõni üksik parandus, et viia OCR tekst inimparandusega identsele kujule. Pikkade tekstide puhul on võimalik, et mitu sõna või mingi arvestatav osa lausest on väga halva OCR kvaliteetidega, kuid see ei peegeldu näite CER ja WER väärtustes. See on võimalik siis, kui näide on piisavalt pikk ja ülejäänud tekst hea kvaliteediga ning siis pole vigade mõju CER ja WER väärtustele piisavalt suur, et need langeksid kõrge müratasemе kategooriatesse. Siiski, kui osa lausest on parandamatu, siis on suur võimalus, et mudel pigem teeb valed parandused, mis toob endaga kaasa CER-i ja WER-i halvenemise.

Meie peenhäälestuste võrdluses on väga pikkade tekstide ehk kõige raskemini parandatavate näidete puhul parimaks Llammas FT-5k. Seevastu Llammas FT-13k on parim ülejäänud pikkustega tekstide parandamisel. Siin sõltub samuti sobiva mudeli valik ülesande iseloomust - kui ei ole võimekust tekstide tükeldamiseks, tuleks eelistada Llammas FT-5k mudelit. Valik Llammas FT-13k mudeli kasuks on mõistlik olukordades, kus väga pikki tekste ei esine. Samuti võib paremate tulemuste saavutamiseks eemaldada ka liiga lühikesed tekstid. Lühikeste tekstide puhul on lisaks eelmainitud vigade suuremale kaalule

veel probleemiks see, et kui paar sõna on täiesti vigased, võib tekstist kaduda oluline kontekst. Sama probleem esineb ka siis, kui tekst on ainult paar sõna pikk ning sellest tulenevalt võib puududa kontekst adekvaatsete paranduste tegemiseks.

5.2.6 Järeldused

Llammas FT-5k ja Llammas FT-13k puhul on paranduste tegemiseks optimaalne andmestik üldise CER väärtustega 9.5% või vähem. WER väärtustega otsest piiri ei ole vaja seada, kuna iga müratasemega oli tulemustes keskmine muutus positiivne, küll aga väga madala WER müratasemega näidete puhul on keskmine muutus pigem väike ja parandustel väiksem mõju. Pikkuse poolest saavad parimaid parandusi tekstid, mille pikkus on vahemikus 25 kuni 124 sõna.

Näidete ideaalseks parandamisel ei hiilga ei meie treenitud- kui ka ChatGPT või DeepSeek V3 mudelid. Parim tulemus on 35 ideaalset parandust 2001-st näitest. Küll aga kui on soov parandada üldiselt näidete kvaliteeti, siis see on keelemudelite abil võimalik. Peenhäälestatud mudelid parandavad või säilitavad CER põhjal kvaliteeti rohkem kui pooltel testandmestiku näidetel. WER tulemuste puhul on see näitaja veelgi parem: 85% - 89%.

Llammase peenhäälestamine ülesandespetsiifilise andmestikuga annab paranduste kvaliteedi poolest selge eelise kommertsmudelite ees, kuid siiski on ChatGPT-4o ja DeepSeek V3 eestikeelsete tekstide parandamise võimekus meie jaoks oodatust kõrgem.

Antud mudelite efektiivne kasutamine tekstide parandamiseks nõuab aga täiendavat kontrollmehhanismi hindamismudeli näol. Kuna raskusastmeteks jaotamisest selgub, et teatavad tekstid peaks parandamisprotsessist elimineerima, aga reaalsuses on suvalisele tekstile lihtsalt peale vaadates raske määrata selle mürataset, siis on vaja luua täiendav mudel, mis sellise hinnangu automaatselt annaks. Samuti on paranduste kontrollimine tüütu ja aeganõudev tegevus. Tööprotsessi kiirendamiseks ning tulemuste parandamiseks on vaja halvimad tulemused automaatselt elimineerida. Selle tarvis pakume lahendust hindamismudeli näol.

6. Hindamismudel

Esmalt räägime probleemidest, mis seonduvad RaRa varasema hindamisskaalaga (peatükk 6.1) ning pakume omaltpoolt välja uue lahenduse peatükis 6.1.2. Seejärel tutvustame hindamismudeli töövoogu peatükis 6.2, peale mida tutvustame lähemalt eraldi kahe erineva hindamismudeli tulemusi peatükkides 6.3 ja 6.4 ning analüüsime neid peatükis 6.5. Järeldusi hindamismudelite ning uue hindamisskaala kohta saab lugeda peatükist 6.6

6.1 Eeltöö

Olemasolevast RaRa hindamismudelist parema ülevaate saamiseks on esialgsed katsetused tehtud Kruusmaa koostatud viibaga ehk toetudes esialgsele subjektiivsele hindamisskaalale (Lisa 4). Selle järgi hindame tekste nii treenimata Llammasega, ChatGPT-4o-ga kui ka ise manuaalselt. Esialgse hindamise läbis ligikaudu 100 teksti - seda eesmärgil, et saada esmane ettekujutus sellest, kui suur on erinevus keelemudelite ja inimeste hinnetel.

Manuaalse hindamise käigus selgus antud metoodika probleemsus. Nimelt, kuna hindamisele lähevad keelemudeli tehtud parandused, mille kvaliteeti määratakse ainult loetavuse ja grammatika poolest, siis on arvestatav võimalus, et inimparandus ja hinnatav tekst on sisu poolest erinevad. OCR tekstide parandamisel on aga äärmiselt oluline säilitada teksti originaalkuju ning senine hindamismeetod sellega ei arvesta. Seetõttu on isegi inimestel tekstidele raske hindeid panna, sest kui inimparandatud teksti ja parandust võrreldes on need sisu poolest erinevad, aga parandus grammatiliselt korrektne ja tekst ise loogiline, siis skaala järgi peab tekst saama ikkagi kõrge hinde.

Seetõttu otsustasime restruktureerida hindamisskaala ning luua kaks erinevat hindamise mudelit. Esimene mudel saab sisendiks OCR teksti ning tagastab selle eeldatava CER-i. Teine mudel saab sisendiks OCR teksti ja paranduse ning tagastab tõenäosuse, et paranduse kvaliteet on parem kui OCR teksti oma.

Metoodika ja töös kasutusel olev hindamisskaala on täpsemalt kirjeldatud peatükis 3.3. Kasutatud viibad leiab lisast 3.

6.1.1 Preemiamudel

Lisaks kahe erineva mudeli hindamismudeli peenhäälestamisele on tehtud katsetus preemiamudeli treenimisega. Preemiamudeli treenimine on analoogne eelistuste suunamisega. Andmestikus on samuti eelistatud ja tagasilükatud väljundid ning mudel peaks treenimise käigus ise kujundama hindamisskaala, kus eelistatud ehk head väljundid saavad kõrgemaid hindeid ja tagasilükatud madalamaid.

Sarnaselt eelistuste suunamisele, ei tööta preemiamudel antud ülesande kontekstis meie jaoks eriti hästi. Treenitud on erinevate andmestikega, kus kohati on kasutusel ainult inimparandused ja OCR tekstid vastavalt eelistatud ja tagasilükatud väljunditena, osades kaasatud ka erinevate mudelite parandused (Llammase ja peenhäälestatud mudelite). Parandused esinevad nii eelistatud kui ka tagasilükatud väljundina, olenevalt CER väärtustest. Treeningandmestikust osa (10%) on kasutusel treenimisel kasutatava kontrollandmestikuna. Tulemused, isegi kui täpsus on kontrollandmestikul kõrged, on pigem juhuslikud. Tulemuste analüüsimisel on raske leida hinnete jaotumisel seaduspärasust ning samuti puudub pealtnäha igasugune seos näiteks teksti CER-i ning saadud hinde vahel. Seetõttu järgnevates tulemustes antud metoodika tulemusi lähemalt ei kajastata ning edasi ei analüüsita.

6.1.2 Hindamisskaala

Hindamisskaala koostamisel on mõlema mudeli korral kasutusel tekstide CER väärtused. Treeningandmestikku hinnete saamiseks tekstide CER-i ennustava mudeli puhul arvutame lihtsalt vigade osakaalu OCR teksti ja inimparandatud teksti vahel, kui hinnatakse OCR teksti, ning paranduse ja inimparandatud teksti vahel, kui hindamisele läheb parandus. Saadud tulemused ümardame täisarvuks, sest komakohtadega täpsus pole selle ülesande puhul oluline. Seeläbi seame igale näitele vastavaks vigade protsendi, aga lihtsuse ja loetavuse huvides esitame protsenti kui hinnet ehk eemaldame tulemuse tabelitest lihtsalt protsendimärgid. Kõige olulisem on meie jaoks see, et antud hinne langeks õigesse suurusjärku, mitte, et vigade osakaal oleks protsendi täpsusega tegelikkusele vastav. Järgnevas tabelis on välja toodud mõned OCR tekstide ning inimparanduste paarid ning hinne, mille OCR tekst saab külge oma CER-ist tulenevalt.

Hinne	OCR tekst	Inimparandus
74	C%#tt nwlxmifli ttwlüfmtfe ajii* liite forK	Eesti wabariigi walitsemise ajutine kord.
9	«ma õpetlasele võimalust teaduslikkudeks uurimuStekS.	anna õpetlasele võimalust teaduslikkudeks uurimusteks.
2	1. aprillist 1920. a. on Tallinna linnr piirides keelatud küpsetada ja müüa igaseltsi koolisid	1. aprillist 1920. a. on Tallinna linna piirides keelatud küpsetada ja müüa igaseltsi kookisid

Joonis 7. *Näited OCR tekstidest ning nende CER hinnetest (madalam on parem)*

Paranduste kvaliteeti hindava mudeli jaoks kasutame hinnete saamiseks CER vähenemist, mis arvutatakse valemiga 3.6. Kui CER väärtused on väga suure tõenäosusega vahemikus 0-100 ning üksikutel juhtudel nendest piiridest väljas, siis CER vähenemise korral on võimalike väärtuste vahemik palju suurem. Seetõttu on skaala koostatud nii, et CER vähenemine 100% annab hindeks 100, vähenemine 0% annab hindeks 50 ning vähenemine -100% või rohkem annab hindeks 0. Kõik ülejäänud väärtused on normaliseeritud selles vahemikus.

Hinne	CER vähenemine	OCR tekst	Parandus	Inimparandus
69	75%	Türgi wägede liitumise seisma puuemiue	Türgi wägede liitumise seisma panemine.	Türgi wägede liikumise seisma panemine
50	0%	Lahtilastud ValUmaa pant* wangide toh a	Lahtilastud Walgamaa pandiwangide toh a	Lahtilastud Baltimaa pant wangide kohta
0	-100%	iääb seäel nädalal imll ära, sest et koolide juhatajatele kvonoe uu»sti - awamise asjus	jääb seekel nädalal ikkagi ära, sest et koolide juhatajatele konwents uuest korrast awamise asjus	jääb sellel nädalal weel ära, sest et koolide juhatajatele koolide uuesti - awamise asjus
0	-160%	£iitlaste ta maštaste föjaftilub.	Ühendatud Wälide wabariigi seadused.	Liitlaste ja wastaste sõjakulud.

Joonis 8. *Näited parandustest ning hinnang sellele, kas ta on OCR tekstist parem (kõrgem on parem)*

Selline lähenemine paranduste hindamisele tagab selle, et parandus ei kalduks liialt algtekstist sisu poolest kõrvale. Kuna CER vähenemine läheb taoliste näidete korral tugevalt miinusesse, siis kajastub see ka hindes ning seetõttu peaks mudel olema suuteline väga halvad parandused üles leidma. Selle skaala üheks probleemiks on see, et kui teksti algne CER on väike ja tekst ise lühike ning paranduse CER läheb ainult natukene halvemaks,

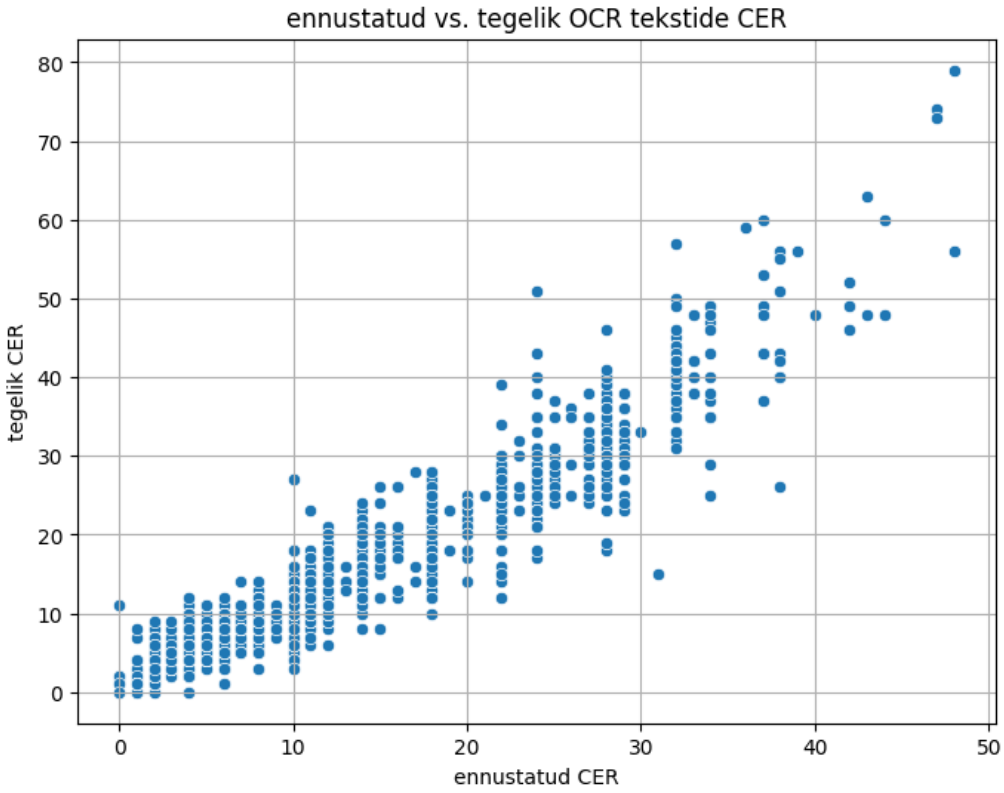
siis CER vähenemine on ikkagi päris tugevas miinuses. Põhjuseks see, et suhtena algse CER väärtuse suhtes, on see muutus suur.

Antud ülesannet lihtsustab võimalus muuta see binaarseks klassifitseerimiseks: parandus kas on või ei ole parem OCR tekstist. Piiriks määrata mingi hinne - vastavast hindest kõrgemad on paremad, madalamad on halvemad. Sellise lähenemise puhul on hea veel paindlikkus - näiteks kui on soov leida parimad parandused, siis saab piiri määrata kõrgemaks, kui on soov leida ainult kõige halvemad parandused, siis võib piir olla madalam.

6.2 Töövoog

Hindamismudelite puhul sarnaneb töövoog suures pildis paranduste mudeli omaga, sest peenhäälestamisel on metoodikas ja koodis vaid minimaalsed muudatused. Viibad, mis on kasutusel hindamismudelite treenimisel, on leitavad lisast 3 (viibad 3.4 ja 3.5).

Vastupidiselt preemiamudelile on peenhäälestatud hindamismudelite tulemused oodatust paremad. Esiteks on vastupidiselt preemiamudeli tulemustele selgelt näha korrelatsioon tekstide kvaliteedi ning hinnete vahel (joonis 9). Samuti on nii meie endi kui ka RaRa jaoks lihtsam ja loogilisem see hinnete jagamise loogika. Tulemusi analüüsides ja hindeid ning tekste kõrvutades on päris lihtsasti aru saada, millised on kummagi mudeli nõrkused.



Joonis 9. *Graafik OCR tekstide ennustatud CER-i kokkulangevus tegeliku CER-iga*

Näiteks, meie esmasel treenimisel saadud paranduste hindamise mudeli selge nõrkus on sellised tekstid, kus parandus näeb küll visuaalselt OCR tekstist parem välja, aga mudel on reaalsuses parandanud sõnu valesti või neid lausa juurde mõelnud. Seetõttu on paljudel valedel parandustel mudeli hinnangul suur potentsiaal olla OCR tekstidest parem. Tulemusi analüüsides joonistub aga viga päris kiiresti välja ning järgmisesse treenimise andmestikku on lisatud juurde palju selliseid näiteid, kus parandus on õigekirja ja loetavuse poolest OCR tekstist parem, kuid paranduse mõttes halb, kuna on sisu poolest erinev.

Mõlema mudeli testimiseks kasutame samuti RaRa testandmestikku 2001 näitega, aga sinna oleme juurde lisanud OCR ja inimparandatud tekstile peenhäälestatud mudelite ennustused. Testitud on FT-5k ja FT-13k mudelite parandustega, kuna need mudelid olid paranduste mudeli tulemustes parimad, aga omavahel võrreldes siiski paranduste poolest natukene erinevad. Testimisel on katsetatud lisaks drastiliselt temperatuuri muutmisega (temperatuur=0.1). Selle tulemusena suurt muutust ei toimunud, tulemused läksid pigem kehvemaks, seetõttu on peenhäälestatud hindamise mudelite testimisel kasutusel samad parameetrid, mis paranduste mudeli puhul.

Treeningandmestik CER-i ennustavale mudelile on sarnane FT-13k mudeli omaga, sest on segu nii erinevatest RaRa tekstidest kui ka sünteetilistest andmetest. Paranduste kvaliteeti hindava mudeli treeningandmestik on aga sarnane eelistuste hindamise andmestikule, sest sinna on lisatud keelemudelite (Llammas treenimata ja treenitud) ennustused.

Sarnaselt paranduste mudeliga on võrdluseks hinnanud tekste ChatGPT-4o ja DeepSeek V3. Viibad on täiendatud ja kohandatud, võrreldes nendega, mida kasutame Llammase peenhäälestamiseks (Lisa 3). Tegu on meie hinnangul parandamisest lihtsama ning konkreetsema ülesandega, kuid on oluline, et kaardistame viibas keelemudelile skaala, mille järgi tõenäosusi või CER-e määrata.

6.3 CER-i ennustamise mudeli tulemused

Peamine viis, kuidas CER-i ennustaja sooritust analüüsime, on arvutades erinevuse tegeliku CER-i ning mudeli ennustatud CER-i vahel absoluutväärtusena.

$$erinevus = abs(CER_{ennustatud} - CER_{tegelik}) \tag{6.1}$$

Nagu eelnevalt mainitud, siis hindamismudelite sooritus on oodatust parem. Esimene positiivne märk on see, et mudel ei tagasta täisarvu asemel teksti või ei jäta välja tühjaks.

Küll aga esineb tühje või tekstilisel kujul vasteid nii ChatGPT kui ka DeepSeek-i testides. Tühjad vastused on eemaldatud, kuna neid esines üksikult, samuti takistavad need objektiivset võrdlemist teiste mudelitega, kuna võrdlemiseks oleks vaja mõelda sinna asemele mingi arvuline vaste, mis võib tulemusi ühes või teises suunas mõjutada ja ei anna ausat ettekujutust, kuna hinne oleks meie poolt määratud.

Testandmestiku OCR tekstide hindamisel on kõige täpsemad peenhäälestatud mudeli vastused. ChatGPT ja DeepSeek on meie peenhäälestatud mudelist selgelt halvemad, kuid on omavahelises võrdluses enam-vähem samal tasemel. Samuti on peenhäälestatud mudelil kordades rohkem selliseid pakkumisi, kus on CER-i hinnanud täpselt õigeks. Võrdlemise huvides lasime OCR tekstide CER-i hinnata ka Llammasel. Llammasega esineb hindamise puhul aga üks suur probleem - pea kõik vastused esitab Llammas valel kujul ehk vahemikus 0-1. Kuna aga osad vastused on esitatud nii nagu peab ehk täisarvuna, siis on korrektsete tulemuste arvutamiseks vaja teha päris palju järeltöötlust.

Õigesse raskusastmesse ennustamine annab kõige parema ettekujutuse sellest, kui täpne on mudel kogu andmestiku lõikes ligikaudse CER-i ennustamisel. Kõige olulisem ongi meie jaoks see, et näiteks madala müratasemega tekstid saaksid CER ennustuse, mis langeb väärtuse poolest samuti madala müratasemе kategooriasse, keskmise müratasemega tekstid vastava ennustuse jne. Raskusastmed on CER müratasemе järgi kaardistatud ja lahti seletatud peatükis 5.2.4.

Järgnevalt selgitame tulemuste tabelis olevate veergude tähendusi.

- Keskmine erinevus – kogu andmestiku ulatuses keskmine erinevus (absoluutväärtus) tegeliku ja ennustatud hinde vahel
- Mediaan – kogu andmestiku ulatuses mediaanerinevus (absoluutväärtus) tegeliku ja ennustatud hinde vahel
- Täpseid pakkumisi – mitmele protsendile näidetele ennustas hindamismudel täpselt õige hinde
- Pakkumine õiges raskusastmes – kui mitu protsenti näidetes langeb ennustatud hinde poolest õigesse CER-i mürakategooriasse

Tabel 9. *OCR tekstide CER-i ennustamise tulemused*

Mudel	**Keskmine erinevus**	**Mediaan-erinevus**	**Täpseid pakkumisi**	**Pakkumine õiges raskusastmes**
ChatGPT-4o	17.40	13.00	2.15%	32.91%

Jätkub...

Jätkub...

Mudel	**Keskmine erinevus**	**Mediaan erinevus**	**Täpseid pakkumisi**	**Pakkumine õiges raskusastmes**
DeepSeek V3	13.62	12.00	1.55%	30.18%
Llammas	13.05	9.00	1.55%	26.55%
Llammas CER-estimator	**2.62**	**2.00**	**19.44%**	**77.34%**

Kui vaadata meie mudeli valesse kategooriasse pakutud näiteid, siis joonistub välja, et mudel kipub hindama tekstide CER-i tegelikkusest natukene madalamaks. Siiski, valesti pakutud näidete puhul on keskmine erinevus ennustatud ning tegeliku CER-i vahel vaid 3.42% ning kuna enim on valepakkumisi keskmises kategoorias, kus müratasemed on juba suhteliselt kõrged, siis pole väikesel erinevusel nii suur mõju.

Tabelis on esindatud need 454 testandmestiku näidet (22.66%), mis Llammas CER-estimator valesse kategooriasse ennustas.

Tabel 10. *Llammas CER-estimator kategooriaeksimused OCR tekstide CER-i ennustamisel*

Tegelik ↓ / Ennustatud →	**Väga madal**	**Madal**	**Keskmine**	**Kõrge**
Väga madal	-	50	0	0
Madal	116	-	39	0
Keskmine	8	170	-	14
Kõrge	0	1	56	-

Samuti laseme CER-i ennustaval mudelil hinnata paranduste CER-i. Peamiselt teeme seda lihtsalt võrdluse mõttes, et näha, kas tulemused on sama positiivsed natukene teise iseloomuga tekstide peal. Samuti on paranduste puhul oluline eripära see, et tekstid on grammatiliselt korrektsed ehk tähevigadeta, kuid võrdluses inimparandatud tekstiga täiesti valed. Seetõttu ei eelda me, et CER-i ennustav mudel väga edukalt nende tegelikku CER taset määrata suudab.

Esimesed ennustused on tehtud Llammas FT-5k mudeli parandustele ning teised Llammas FT-13k parandustele. Seda sellel lihtsal põhjusel, et mudelite paranduste kvaliteet on kohati erinev ning erinevate näidete peal testides saame parema ülevaate CER-i hindamise mudeli tegelikust võimekusest.

Llammasel pole lastud ennustusi hinnata, sest juba OCR-i ennustamisel on näha, et kommertsmudelitega on see täpsuse poolest umbes samal tasemel, aga sellised tulemused on saadud alles peale mitmeosalist järeltöötlust. Seetõttu ei näe me, et Llammas võiks paranduste ennustamisel märkimisväärseid tulemusi saavutada.

Tabel 11. *Paranduste CER-i ennustamise tulemused*

Mudel	**Keskmine erinevus**	**Mediaan erinevus**	**Täpseid pakkumisi**	**Pakkumine õiges raskusastmes**
Llammas FT-5k ennustused				
ChatGPT-4o	13.22	8.00	4.75%	31.50%
DeepSeek V3	12.95	9.00	4.50%	31.13%
Llammas CER-estimator	**10.45**	**6.00**	**7.15%**	**34.98%**
Llammas FT-13k ennustused				
ChatGPT-4o	12.57	8.00	4,45%	33.35%
DeepSeek V3	11.95	8.00	3.75%	31.20%
Llammas CER-estimator	**10.79**	**6.00**	**7.30%**	**34.18%**

Nagu näha, siis langeb peenhäälestatud mudeli täpsus paranduste ennustamisel suurel määral. See tuleneb suures pildis juba eelnevalt mainitud probleemist, kus parandus on grammatiliselt küll korrektne, aga sisu poolest inimparandatud tekstist erinev. Peenhäälestatud Llammas on küll kõigis kategooriates jätkuvalt parim, kuid vahed täpsuses võrdluses teiste mudelitega on kordades väiksemad kui OCR-i CER-i ennustamisel. Kategooriate kaupa valesid pakkumisi analüüsides selgub taaskord, et mudel hindab tekstide CER-i tegelikkusest madalamaks. Kõige rohkem eksimusi on keskmise- ja madala müratasemе kategooriates, sest mudel ennustab tekstide CER-i vastavalt siis madalaks või väga madalaks.

Kui lisada eksimispuhver, mis lubab mudelil ennustada kuni kolme protsendi võrra mööda, siis tõuseb õigesse kategooriasse pakutud näidete arv FT-5k mudeli ennustuste puhul 38.28% peale ja FT-13k vastavalt 40.13% peale. Kui tõsta puhver kuueks, on täpsus vastavalt 54.42% ning 55.27%. Seega on päris suur hulk näiteid, mille puhul ennustab mudel mõne protsendiga mööda. Siiski on meie jaoks täpsus kategooriatesse ennustamisel liiga madal ning seetõttu jääb paranduste CER-i ennustamine vaid eksperimendiks ning mudelite koostöös kasutamiseks ei pea me mudeli sooritust piisavalt veenvaks.

6.4 Paranduste hindamise mudeli tulemused

Sarnaselt paranduste CER-i ennustamisega valmistab mudelitele raskusi paranduste kvaliteedi hindamine võrdluses OCR tekstiga. Parimad tulemused on taaskord meie peenhäälestatud mudelil. Meie mudelile tekitavad peamiselt probleeme sellised parandused, mille CER vähenemine on nulli lähedal, sest üldjuhul hindab nende kvaliteeti natukene liiga kõrgeks. ChatGPT ja DeepSeek puhul võib lugeda katsed suures osas ebaõnnestunuks. Arvestades erinevusi ei tundu tulemused õigesti klassifitseerimisel isegi liiga halvad, aga tegelikkuses ennustab mudel lihtsalt juhuslikult ning sellest ka 50% lähedal olev täpsus.

Tulemuste puhul on mudeli täpsuse osas kõige olulisem näitaja õige klassifikatsioon. Õige klassifikatsioon tähendab siinkohal seda, et mudel suutis anda parandusele sellise hinde, mis jääb tegeliku paranduse kvaliteediga meie seatud künnisest samale poole. Künniseks seadsime 51% ehk tõenäosuse, et parandus on parem OCR tekstist.

Tabel 12. *Paranduste hindamise tulemused*

Mudel	**Keskmine erinevus**	**Mediaan erinevus**	**Täpseid pakkumisi**	**Õige klassifikatsioon**
Llammas FT-5k ennustused				
ChatGPT-4o	39.80	37.00	1.00%	46.03%
DeepSeek V3	39.74	37.00	0.65%	46.83%
Llammas prediction-quality-estimator	**13.32**	**7.00**	**5.35%**	**60.32%**
Llammas FT-13k ennustused				
ChatGPT-4o	40.37	38.00	1.00%	49.63%
DeepSeek V3	40.13	38.00	0.75%	50.12%
Llammas prediction-quality-estimator	**12.20**	**7.00**	**6.40%**	**61.62%**

Samamoodi nagu paranduste CER-i hindamisel, on ka siin probleemiks sellised tekstid, mis on grammatika poolest head, aga sisult tegelikult valed. Isegi kui võrdluseks on olemas viibas OCR tekst, mille põhjal parandus tehti, siis kohati ei ole võimalik aru saada, kas mudeli pakutud paranduse peaks lugema õigeks või valeks.

Ennustatud hinne	Tegelik hinne	OCR tekst	Parandus	Inimparandus
100	35	NSuptäamwo lõjawao varustawllo ja riiglkaltlo asjus.	Riigikontrolööri lõjawangide varustamise ja riigikontrolli asjus.	Nõupidamine sõjawäe warustamise ja riigikaitse asjus.

Joonis 10. *Näide, mille puhul parandus on õigekirja poolest korrektne, aga sisult vale*

6.5 Tulemuste analüüs

Peenhäälestamine on väga edukas CER-i ennustamisel just OCR tekstide jaoks. See tuli natukene üllatusena, kuid samas on see keelemudelile piisavalt lihtne ülesanne, mida täita. Oluline saavutus on veel märgatav vahe täpsuses meie peenhäälestatud mudeli ning kommertsmudelite ja treenimata Llammase vahel. Kuigi eeldasime, et ChatGPT ja DeepSeek on võimelised tekstis tuvastama tähemärgivigu ning seejärel protsendi tagastama, siis oletame, et suurt rolli mängivad andmestiku eripärad, mistõttu treenimine parandab oluliselt tulemusi. CER-i ennustamine paranduste puhul nii täpne ei ole ning seda kohe algusest peale ka eeldasime, mistõttu nägime vajadust veel teise hindamismudeli järele, mis kõrvutaks paranduse algse OCR tekstiga ning aitab seeläbi tuvastada algsest tekstist sisu poolest kõrvale kalduvad parandused.

Vähem edukaks osutus paranduste kvaliteedi hindamine. Seda küll mingil määral eeldasime, kuid ootus oli, et vähemalt õigele poolele lävendit ennustamise täpsus oleks parem kui 60%. Arvame, et tegu on keerulisema ülesandega kui esmalt eeldasime. Seda näiteks põhjusel, et mõne teksti OCR müratase on nii kõrge, et raske on määrata, kas keelemudeli parandus on õige või mitte ilma algteksti nägemata. Lisaks on andmestikus näiteid, kus parandus on väga väiksel määral halvem OCR tekstist, kuid saab näiteks hindeks 51. Eksimuse poolest see nii suur vahe ei ole, kuid klassifitseerimise mõttes on vale.

Paranduste kvaliteedi hindamisel oli oluline eesmärk, et halvimad parandused, kus selgelt mudel on algtekstist täiesti mööda parandanud, sõnu kordama jäänud või näiteks lause otsa lisanud, saaksid halva hinde. Suures pildis saab aga mudeli soorituse selliste näidete ennustamisel lugeda edukaks. Kui filtreerida välja 20 kõige negatiivsema CER muutusega parandust lävendi mõlemalt poolt ja liita CER muutused kokku, siis selgub, et summa on rohkem kui kaks korda negatiivsem lävendi all, mis näitab, et kõige kehvemad parandused on mudel õigesti tuvastanud. Samuti, kui võtta vastupidi just 20 kõige parema CER muutusega teksti, siis on tulemus rohkem kui kolm korda kõrgem lävendist ülespoole jäänud hulgal.

6.6 Järeldused

Kuigi mudelid on treenitud peamiselt koostööd silmas pidades, et tõsta paranduste kvaliteeti, siis on mudelid ühtlasi kasutatavad eraldiseisvatena. Näiteks, CER-i hindavat mudelit saab RaRa kasutada selleks, et hinnata DIGAR-is olevate kontrollimata OCR tekstide kvaliteeti. Kuna mudeli täpsus OCR tekstide puhul on väga hea, siis saab see olema neile oluline tööriist ka väljapool tekstide parandamist. Näiteks väga kehva CER kvaliteediga tekstide korral saab RaRa otsustada, et kas on mõttekam teksti uuesti skaneerida, kui parandada. Samamoodi on võimalik CER-i hindaja abil kiiresti üles leida tekstid, mille kvaliteet on väga hea või ideaalne ning need hoopis parandamisest ja uuesti skaneerimisest elimineerida.

Lisaks ei ole teada, millist teksti milline OCR mudel on töödelnud ning kuna mudelite töökvaliteet on kõikuv, siis on ka digiteeritud tekstide CER tasemed varieeruvad. Kuna tekste on väga palju, siis oleks selle väljaselgitamine manuaalselt ülimalt ajamahukas. Seetõttu on CER ennustajal oluline roll ka tekstide kategoriseerimisel. Mudeli abil on võimalik suhteliselt täpselt ennustada mingi väljaande CER-i ning sellest lähtuvalt otsustada, mida väljaandega ette võtta. Näiteks, suure või väga suure müratasemega tekste on ehk parem suunata kohe inimestele parandamiseks, sest keelemudel ei suuda suure tõenäosusega piisavalt täpset parandust pakkuda. Teine võimalus on lasta taoliseid tekste uuesti OCR tehnoloogia abil töödelda või olenevalt PDF kvaliteedist uuesti skaneerida.

Hindamismudelite eesmärk on ühtlasi toetada paranduste mudelit ning seeläbi tõsta üldist paranduste kvaliteeti. Täpsemalt on kirjeldatud mudelite koostöö töövoog, tulemused ning analüüs ja järeldused peatükis 7.

7. Mudelite koostöö

Paranduste mudeli puhul on üheks suurimaks kitsaskohaks sõnade väljamõtlemine ja juurde lisamine. Sellised parandused on just selle poolest kõige kehvemad, et algteksti nägemata on ka inimsilmal raske aru saada, kas teksti sisu on õige või on mudel lisanud raskesti parandatavatesse kohtadesse omaloomingut. See aga takistab jällegi paranduste töövoogu, sest kontrollija peab käsitsi kõik parandused üle vaatama. Probleemi leevendamiseks pakume välja prototüübi mudelite koostöö töövoost, mis kasutab kõiki meie poolt peenhäälestatud mudelit.

Esmalt kasutame OCR tekstide CER-i hindavat mudelit, mis filtreerib välja kahte sorti tekstid: need, mille kvaliteet on väga hea (ennustatav CER 0-2%) või sellised tekstid, mille kvaliteet on väga halb (ennustatav CER üle 45%). Heade tekstide välja filtreerimise põhjuseks on see, et selliste näidete puhul läheb pea alati kvaliteet natukene halvemaks, kuna väga hea kvaliteediga näiteid on vähe ning mudel on treenitud suures osas näidete peal, mille CER on kõrgem. Samuti on sellises suurusjärgus CER-iga näidete vigade arv väike ning teksti kvaliteeti suuremahuliselt ei mõjuta. Väga halvad tekstid eemaldame samal põhjusel - mudel ei saa nii vigaste tekstide parandamisega hästi hakkama, kuna tekstis võib esineda palju moonutatud sõnu, mille parandamine algse dokumendita võimalik ei ole. Sellisel juhul eemaldame need üldse andmestikust, kuna antud tekstid vajaksid inimparandusi või uuesti OCR tehnoloogiaga töötlemist.

Teise sammuna küsime parandused allesjäänud tekstidele. Tulemuste tabelis on kajastatud eraldi nii FT-13k mudeli kui ka FT-5k mudeliga saadud tulemused. Parandusi laseme hinnata paranduste kvaliteeti hindaval mudelil. Mudelite koostööl on paranduse hinde lävendiks 40 punkti, et välja jääksid vaid kõige kehvemad parandused. Kuna meie mudeli sooritus nii veenev ei olnud keskpäraste paranduste puhul, siis seetõttu prioritiseerime koostööl ainult halvimate leidmist.

See jaotab jällegi andmestiku kaheks: all- ning üleval pool lävendit olevad tekstid. Lävendi alla jäänud tekstid saadame kohe uuesti paranduste ringile. Üleval pool lävendit olevatele tekstidele teeme pikkusekontrolli - kui parandus on OCR tekstist 5% pikem või lühem, siis läheb ka see uuele paranduste ringile. Pikkusekontroll annab veel lisavõimaluse eemaldada selliseid parandusi, kus OCR tekst ja parandus ei klapi sisu poolest.

Uuesti parandamisele lähevad seega sellised näited, kus hindamismudel tuvastab kohe

kehva kvaliteedi ning sellised parandused, mis ei klapi OCR tekstiga pikkuse poolest ehk parandus on lisanud lauseid, sõnu või neid välja jätnud. Parandusi teostab ikka sama mudel, mis esimesel ringil, vastavalt siis FT-13k või FT-5k, aga muudame teise ringi jaoks kolme treeningparameetrit: temperatuur, top_p ning top_k. Nagu paranduste mudeli peatükis mainisime, siis on meie töös need vaikimisi vastavalt 0.7, 0.1 ning 40. Nende parameetrite tõstmine suurendab aga mudeli loomingulisust ning lubab vastustel rohkem varieeruda. Seetõttu tõstame teisel parandusteringil nende väärtused vastavalt 0.8, 0.95 ning 50 peale.

Peale teise ringi parandusi hindab paranduste hindamismudel tekste uuesti ning filtreerime tekste sama loogika ja lävendi põhjal, mis esimeselgi korral. Samuti läbivad hea kvaliteediga tekstid uuesti pikkusekontrolli. Seekord on aga erinevus selles, et kui mõni parandus jääb alla 40 punkti lävendi, siis ei saada me seda kolmandale parandusringile, vaid asendame algse OCR tekstiga. Seda seetõttu, et meie hinnangul on selliste näidete korral parandajal lihtsam parandada algset OCR teksti, mitte paranduste mudelist tulnud teksti, milles näiliselt pole vigu, aga teksti sisu on vale. Küll aga on võimalik RaRal tulevikus ise valida, mitu korda tekste parandatakse.

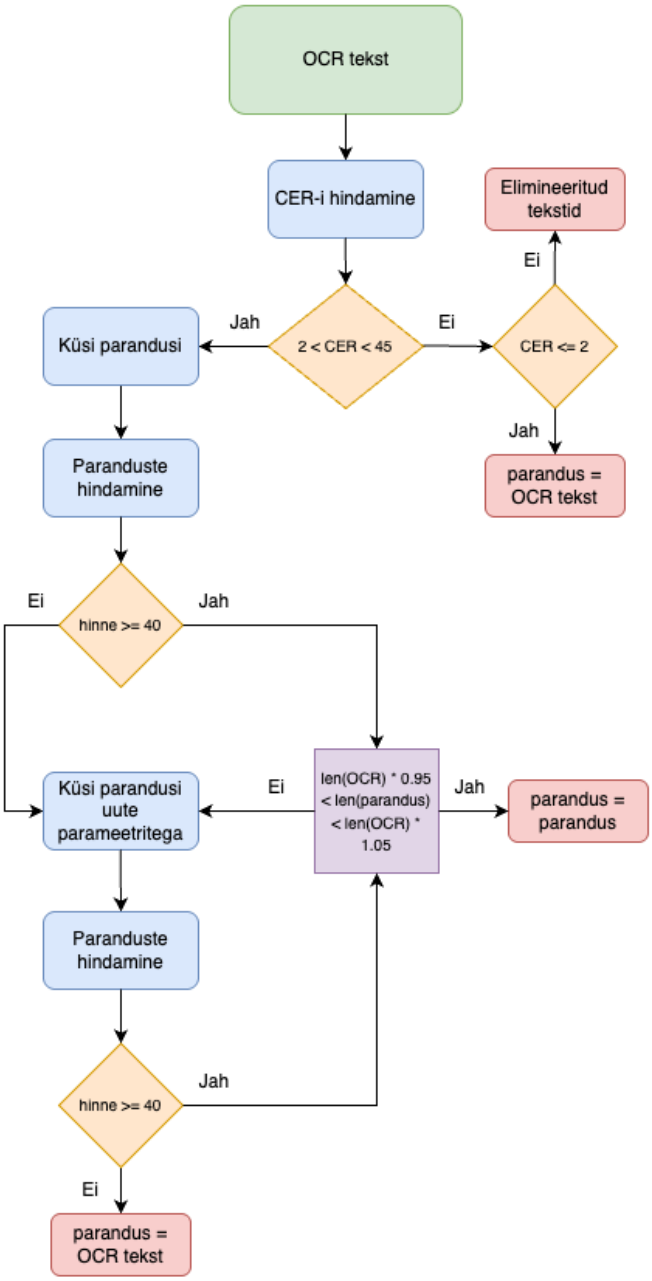
Lõpuks paneme ühte faili kokku esimesest sammust saadud väga head OCR tekstid, esimese paranduste ringi head tekstid, teise paranduste ringi head tekstid ning viimaks tekstid, mille puhul jääb paranduse asemele OCR tekst. Eraldi tagastame faili, mis sisaldab esimeses ringis välja filtreeritud tekste, kus OCR väärtus on üle 45%.

Mudelite koostöö töövoog on kokkuvõtvalt kujutatud joonisel 11.

7.1 Tulemused

Koostöö tulemusena on CER muutus positiivne, mis tähendab, et andmestiku kvaliteet lõppkokkuvõttes paraneb, kuid seda küll väiksel määral. Siiski, kui vaadata paranenud või sama kvaliteedi peale jäänud näidete osakaalu, siis on tulemused head. Kui parima paranemiste osakaalu ainult paranduste mudelit kasutades sai FT-13k mudel, parandades või jättes samaks seejuures 55.07% tekstidest, siis sama mudeli parandusi kasutades tõuseb koostööl paranemiste osakaal pea 20%. FT-5k mudeli puhul on paranemine ligikaudu 15%.

Mediaani puhul on sulgudes eraldi välja toodud ka esinemissageduse poolest teisel kohal asetsev väärtus, sest kõrge nullide sagedus tuleb suures pildis asjaolust, et kõige kehvemad parandused asendatakse lõpuks OCR tekstidega ning seetõttu on ka CER muutus sellistel näidetel null. Selliseid näiteid on FT-13k mudelil 270 ning FT-5k mudelil 249 ehk natuke alla viieteistkümne protsendi. Lisaks on andmestikust eemaldatud ning eraldi faili tõstetud



Joonis 11. Koostöö protsessi skeem

13 teksti, mille ennustatav OCR teksti CER on üle 45 protsendi. Kuna selliste näidete puhul on peaaegu võimatu ilma originaalteksti nägemata aru saada, milline tekst tegelikult olema peaks. Selliste tekstidega saab lõpuks RaRa ise otsustada, mis peale hakata.

Nagu ka parimast ja halvimast muutusest on näha, siis on mõlema mudeli puhul halvim muutuse väärtus palju suurem positiivse muutuse omast. See on ka suur põhjus miks keskmine muutus on üle andmestiku niivõrd väike, kuigi suurem osa näiteid paraneb. Vaatamata sellele, et meie hindamismudelid suudavad koostöös leida üles halvimad parandused (varem oli halvim muutus FT-5k mudelil, -80.24% ja FT-13k mudelil -128.26%), kehtib siiski üldreegel: halbade paranduste kahjutegur on sageli kordades suurem heade paranduste positiivsest mõjust. Seetõttu langetavad üksikud sissejäänud halvad parandused oluliselt keskmist väärtust. Samuti on Llammas FT-5k mudeli paranduste puhul ära kadunud eelnev parim muutus. Samuti peame ära mainima, et järgnevas tabelis olevad tulemused ei ole üks-ühele võrreldavad tabel 2 tulemustega, sest koostöö tulemustest on eemaldatud 13 "parandamatut" ehk väga kõrge CER väärtusega OCR teksti.

Tabel 13. *CER muutuste statistika mudelite koostööl*

Mudelid	**Keskmine**	**Mediaan**	**Parim muutus**	**Halvim muutus**	**Muutus >= 0**	**Ideaalne**
Llammas FT-5k ja hindamine FT	0.19%	0.00% (1.00%)	20.00%	-39.00%	67.52%	1.00%
Llammas FT-13k ja hindamine FT	0.66%	0.00% (1.00%)	32.08%	-46.00%	73.91%	1.95%

7.1.1 Näite puhul tsükkel läbi mudelite

Järgnevalt näitame koostööstüklit ühe näite puhul.

Tabel 14. *Tekstinäide testandmestikust enne töötlemist*

Inimkontrollitud tekst	**OCR tekst**	**Ennustatud CER**	**Tegelik CER**
Wikat wigastas wäetikest ja kooliõpilast.	MiKat wigasias wiietikest jn kooliõpilast.	12%	15%

Kuna OCR teksti puhul pole CER ei väga halb ega väga hea, siis läheb tekst parandamisele.

Tabel 15. *Tekstinäide testandmestikust enne töötlemist*

Parandus	**Ennustatud hinne parandusele**	**Tegelik hinne**
MiMat wigastas wieteist wiietkest ja kooliõpilast.	38	0

Kuna ennustatud hinde poolest jääb parandus täpselt alla lävendit, siis saadetakse see otse uuele parandusele.

Tabel 16. *Tekstinäide testandmestikust enne töötlemist*

Uus parandus	**Uus ennustatud hinne parandusele**	**Tegelik hinne**
Nikat wigastas wieteistkümmet ju kooliõpilast.	50	29

Seekord sai parandus kõrgema hinde ja seetõttu läbib see ka teise ringi paranduste hindamise kontrolli ning parandus tagastatakse kasutajale sellisena, nagu mudel seda parandas.

7.2 Analüüs

Peamiseks kitsaskohaks peame koostöö puhul seda, et kuigi parimal juhul paraneb või jääb samaks üle 73% näidetest, siis peab tulemuste osas olema ikkagi skeptiline. Kuna ajalooliste tekstide digiteerimisel ja parandamisel on väga oluline säilitada teksti originaalkuju, siis peab eriti just madalama hinde saanud paranduste puhul olema eriti tähelepanelik, et mudel ei oleks midagi juurde lisanud või valeks muutnud. Näiteks, FT-13k halvim parandus, mis jõudis lõplikku faili välja, on täiesti erineva sisuga OCR tekstist, aga grammatiliselt täiesti korrektne. See aga jätab ikkagi vajaduse inimkontrolli järele.

Positiivse poole pealt aga kõige halvemate paranduste eemaldamine suures pildis õnnestus ning juba see vähendab kontrollijate töömahtu. Samuti oleme õnnelikud selle üle, et isegi vaatamata mõningatele kehvadele sissejäänud parandustele on keskmine CER muutus üle andmestiku ikkagi positiivne. Lisaks pakub CER-i hindamise mudel olulist vabadust RaRale tulevikus koostöö osas otsustada, milliseid parandusi keelemudelile üldse saata. Meie poolt pakutud vahemikud ei ole fiktseeritud ning on vabalt muudetavad vastavalt and-

mestiku eripäradele ja kasutaja soovidele. Samuti on võimalik kasutada teisel paranduste ringil alternatiivseid parameetreid.

Arvame, et mudeleid on koostöös võimalik hea efektiivsusega kasutada, kuid alguses tuleb tulemustesse suhtuda ettevaatlikult. Kuna teste oleme teinud ainult RaRa testandmestiku peal, siis ei tea me päris täpselt, mis oleksid täpsusnäitajad kogu DIGAR-i puhul. Seetõttu edukas koostöö implementeerimine nõuaks suuremas mahus testimist ja analüüsi ning antud lõputöö käigus meie poolt pakutud tööprotsessi tuleks käsitleda eelkõige prototüübina. Lisaks vajaks parima koostöö leidmine erinevate parameetritega laiaulatuslikku testimist, mida antud lõputöö raames teostatud ei ole, kuna võimalikke muudetavaid parameetreid on palju. Näiteks on võimalik uuesti parandada tekste ühe korra asemel lõpmatul hulgal.

8. Järeldused

8.1 Kokkuvõte tulemustest

Töö esimeseks väljundiks on RaRa testandmestiku manuaalne kontrollimine ning andmestikus esinevate probleemide kaardistamine. Testandmestiku 2001 näidet on kontrollitud, kuid treeningandmestik jäi kontrollimata ning andmestiku puhul säilivad esinenud probleemid (erineva pikkusega tekstid, poolikud laused, vead jne). Sünteetiliste tekstide genereerimine oli edukas andmestiku laiendamises, samuti aitas vähesel määral tõsta paranduste kvaliteeti ning ühtlustada andmestiku eripärasid.

OCR tekstide automaatse järeltöötluse tulemusena valmis kaks peenhäälestatud keelemudelit. Paranduste kvaliteet varieerub üpriski suurel määral - esineb tekste, kus algne OCR kvaliteet on väga halb, kuid parandus üllatavalt hea, seevastu on tekste, kus OCR vead on tüüpvead ning parandamine peaks olema lihtne, kuid mudel edukalt seda teha ei suuda. Probleem on samuti see, et parameetrite muutmist (eelkõige temperatuur) sai üsna vähesel määral katsetatud. Praegu on temperatuur (0.7) selline, et sama teksti puhul võivad parandused väga palju erineda.

Samuti on teostatud detailanalüüs raskusastmete kaupa, mis võtavad arvesse algset CER ja WER taset ning OCR tekstide pikkust. Sealt selgub, et eelkõige CER puhul väga kõrge müratasemega muutub parandamine keelemudelitele järjest raskemaks. Pikkade ja väga lühikeste tekstide puhul keskmine paranduse kvaliteet samuti langeb. Kommertsmudelitega võrreldes aga oleme nähtavalt edukamad, eriti suur vahe on WER väärtuste poolest. Paremad tulemused treenimisel võiksid sellest tulenevalt olla saavutatavad siis, kui treeningandmestik oleks parema CER kvaliteediga, tekstid ühtlasema pikkusega ning andmestikus oleks tekstid avaldatud lühema ajavahemiku jooksul.

Tekstide kvaliteedi hindamine sai töö käigus ümber mõtestatud ning jaotatud kahte erinevasse etappi. OCR tekstide kvaliteedi hindamise puhul on eelkõige olulisim näitaja CER, mistõttu on töö käigus treenitud CER-i ennustav mudel, mille tasemega oleme väga rahul. Teisena on oluline paranduste hindamiseks võrdlus algse OCR teksti ning parandust vahel, seetõttu on treenitud mudel, mis hindab tõenäosust, et parandus on parem kui algne OCR tekst. Selle tulemused on oodatust kehvemad, kuid oluline on siinkohal välja tuua, et suure hulga tekstide puhul osutub selline hindamine ikkagi päris keeruliseks ülesandeks, kui ei ole võimalust alusteksti vaatama minna. Seda just sõnade puhul, mis esinevad OCR

tekstis moonutatud kujul ning paranduse puhul ei saa kindlalt määrata, kas parandus on vale või mitte. Seetõttu selle ülesande puhul siiski ei näe me hetkel võimalust keelemudelit heal tasemel kasutada kõikide paranduste hindamiseks. Küll aga on meie lahendus edukas, et leida üles väga head või vastupidi väga halvad parandused ning sellele rõhusime ka mudelite koostöö implementeerimisel.

Suurim ebaõnnestumine töö käigus on nii eelistuste suunamise (DPO) kui ka preemiamudeli treenimise osas. Mõlema tulemused on väga kehvad ning seetõttu suures mahus tulemustes ei kajastu. Eeldame, et üks osa probleemist võib olla, et nende metoodikate jaoks ei olnud meie andmestik piisavalt kvaliteetne. Kuna metoodikate puhul on kasutusel eelistatud ja tagasilükatud väljundid, kuid keelemudel nende valimiseks mingit põhjendust ei saa, siis paranduste ja hindamise puhul võib seda olla väga raske leida, kuna ülesanded on komplekssed ning käsitlevad väga paljusid erinevaid nüansse. Samuti kasutatakse eelistuste suunamist rohkem loominguliste vastuste kujundamiseks, meie ülesannete puhul on aga alati üks konkreetne vastus, mistõttu see metoodika ei pruugigi antud ülesande lahendamiseks sobida. Lisaks on võimalik, et Llama-2 põhjal treenitud Llammas ei ole sobilik nende metoodikate jaoks, kuna see on üsna väheste parameetritega mudel ning samuti natuke vanem kui mudelid, millele neid metoodikaid peamiselt rakendatakse.

Võrdluses kommertsmudelitega on selgelt meie treenitud mudelid paremad. Väiksem erinevus on paranduste puhul, kus enim erineme positiivse muutusega näidete osakaalu ning WER tulemuste poolest. Suurim erinevus on CER-i ennustamise puhul, mille puhul meie mudeli tase on väga hea, kuid kommertsmudelid üldse edukad ei ole. Samuti oleme edukamad paranduste hindamisel. Oleme ka rahul sellega, et peenhäälestamise tulemusena paranes ülesannete lahendamise võimekus Llammase mudelil märgatavalt. See annab lootust, et eestikeelne mudel nagu Llammas, võiks tulevikus olla võimeline lahendama paljusid eriilmelisi ülesandeid. Vahe treenimata mudelitega tekstiparanduste puhul süveneb veelgi kui integreerida töö tulemusena valminud mudelitevaheline koostöö.

Koostöös hindamise mudelitega on tekstide parandamine edukam kui paranduste mudeli eraldi kasutamine. Koostöö elimineerib halvimad parandused ning suurendab sama- või paranenud kvaliteediga näidete osakaalu märgatavalt. Hindame lahendust rahuldaval tasemel edukaks, kuna suuresti suudetakse koostöö käigus halvimad parandused elimineerida. Puudujääk on väikese testandmestikuga testimine ning parameetritega vähene katsetamine. Lisaks annab koostöö suure paindlikkuse vastavalt oma andmestiku ja soovide järgi erinevaid kriteeriumeid ja parameetreid muuta, et tagada soovitud tulemus.

8.2 Meie töö kitsaskohad

Esmane oluline puudus antud töö puhul on andmestik. Selle juures esineb mitu probleemi: treeningandmestik on kontrollimata ja vigane, andmestiku üldine müratase sarnaste töödega võrreldes kõrgem, andmestiku ajaline ulatus on suur ning andmestikus on tekstid varieeruva pikkusega. Oluline on veel märkida, et RaRa andmestik antud töös moodustab vaid väga väikese osa kogu DIGAR-i andmestikust, mistõttu on võimalik, et teatud levinud veakohad ja probleemid on töö käigus kaardistamata ning vajaksid lähemat uurimist.

Samuti on antud töö raames üleüldine ressursiline piirang. Näiteks ei olnud võimalik katsetada DeepSeek treenimisega, kuna eraldi andmestiku koostamine on ajamahukas ning arvutusvõimekus suuremate mudelite peenhäälestamiseks puudub. Samuti ei olnud samadel põhjustel võimalik teha sügavamat analüüsi parameetrite mõjule antud ülesannete puhul. Parameetritega on katsetatud vähesel määral ning tervikpilt puudub, kas mõne parameetri muutmisega oleks näiteks võimalik märkimisväärselt keskmise paranduse taset tõsta. Hindamise puhul on ühekordne temperatuuriga katsetus tehtud, kuid sellest samuti laialdasi järeldusi teha võimalik ei ole.

Ühtlasi on raske öelda, kas lõputöö tulemusi on enim mõjutanud meie metoodikad, andmestikud või hoopis eesti keele olemus, kuna selle jaoks oleks vajalik põhjalikum ja laialdasem analüüs erinevate mudelite ning andmestike lõikes.

8.3 Võimalikud edasiarendused

Esimese võimaliku edasiarendusega oleks viia läbi samalaadi katsetused kontrollitud treeningandmestikuga. Andmestik tuleks ühtlustada kvaliteedi ja teksti pikkuste poolest. See aitab selgitada, kas konkreetsema ja kvaliteetsema andmestikuga on võimalik sama metoodikaga saavutada paremaid tulemusi. Lisaks on võimalik jooksutada meie poolt treenitud mudeleid suuremas mahus DIGAR-is esineva andmestiku peal, et veel paremini hinnata tugevusi ja nõrkusi.

Ühtlasi on võimalik treenida alternatiivseid keelemudelid samu ülesandeid täitma, et kaardistada ära, kas ülesanne ise ongi keeruline, või on mingil määral see seotud ka konkreetsete keelemudelite võimekusega. Näiteks võiks proovida treenida DeepSeek V3 ja R1 mudeleid, samuti 2025. aasta kevadel ilmunud Gemma 3 mudelit. Lisaks, kui Eestis peaks ilmuma uuem eestikeelne mudel kui Llammas, siis samu metoodikaid ja andmestikku saab rakendada ka uue mudeli treenimisel.

Samuti on võimalik kasutada töö tulemusena saavutatud hindamismudeleid treeningprotsessis tagasisidestamiseks või sobivate väljundite valimiseks. Seda eelkõige näiteks eelistuste suunamise puhul, kus just kvaliteet mängib eelistatud ja tagasilükatud väljundite puhul kõige suuremat rolli.

Uurida on veel vaja laias mahus treenitud mudelitele parameetrite muutmise mõju ning kas oleks võimalik sooritust parandada näiteks puhtalt ainult mudeli parameetreid muutes.

Koostöö puhul vajaks meie poolt pakutud lahendus laialdasemat analüüsi ja testimist. Antud töös pakume koostööd prototüübina, et anda ülevaade, kuidas treenitud mudelid koos saaksid töötada ning mis oleks mudelite koostöö potentsiaal. Siiski ei ole antud töö mahus laialdaselt uuritud teise ringi paranduste parameetrite vahetamist, rohkem kui üks kord uuesti paranduste genereerimist ning testimist suurema andmestikuga.

RaRa pole ainuke asutus, kus OCR tehnoloogiat kasutatakse ning ajaloolised tekstid ainuke variatsioon võimalikest sisenditest. Meie töö käigus kasutatud metoodikaid ja tulemusi on võimalik rakendada ka teistes valdkondades ja eriilmelistele eestikeelsetele tekstidele teatavate muudatustega.

8.4 Hinnang tööle

Arvestades, et antud töö on esimene selline eestikeelse andmestiku ja eestikeelse keelemudeliga, ning meie poolt koostatud hindamise metoodika ainulaadne, siis hindame oma tööd oluliseks, vajalikuks ning oleme oma tulemustega rahul. Arvestades, et OCR järeltöötlus on lahendamata probleem ka mujal maailmas ning OCR tekstide kvaliteedi hindamist on samuti üsna vähesel määral uuritud, on antud töö käigus saavutatud tulemused ja tehtud järeldused oluline panus eestikeelsete keelemudelite arengusse, võimekuse kaardistamisse ning konkreetse valdkonna kaardistamisse eestikeelsete mudelite ja andmestikega.

Kuigi töö tulemusena OCR tekstide parandamise ja kvaliteedi hindamise automatiseerimine keelemudelitega täies mahus veel võimalik ei ole, siis mudelite koostööna saavutatud tulemused annavad lootust tuleviku osas. Samuti on töö käigus selgunud informatsioon ning tehtud analüüs olulised RaRale ning toetavad nende asutuses eesti kirjakeele säilitamise ja uurimise protsesse.

9. Kokkuvõte

Lõputöö eesmärk on teostada RaRa jaoks OCR tekstide automaatne parandamine ja OCR tekstide ning paranduste kvaliteedi hindamine kasutades suuri keelemudeleid. Töös on metoodikatena kasutusel peenhäälestamine, eelistuste suunamine ning preemiamudeli treenimine. Kasutusel olev andmestik koosneb RaRa poolt koostatud andmestikust, kuhu on kaasatud ühisloome inimparandused, ning sünteetiliste vigadega andmestikust, kus on kasutatud nii RaRa inimparandusi kui ka ilukirjandusteoseid.

Tulemusena valmis neli keelemudelit, mis on treenitud eestikeelse mudeli Llammas baasil: kaks neist on tekstiparandusteks ja kaks tekstide kvaliteedi hindamiseks. Võrdlusena on nii parandusi kui ka hindeid genereerinud ChatGPT-4o ja DeepSeek V3, treenimise edukuse hindamiseks ka Llammas vestlusrobot. OCR tekstide parandamine on keeruline ülesanne ning keelemudelid seda veel edukalt teha ei suuda. Hindamise puhul suudab meie mudel OCR teksti CER-i ennustada väga heal tasemel, paranduste kvaliteedi hindamine on siiski veel keeruline, kuid mudel on edukas väga heade ja halbade tekstide leidmises. Kahjuks ei õnnestunud eelistuste suunamise ja preemiamudeli metoodikad.

Paranduste töövoo automatiseerimiseks on integreeritud mudelite koostöö prototüüp, mille tulemusena andmestiku keskmine paranduste kvaliteet paraneb. Lisaks jätab koostöö palju ruumi varieerimiseks tulevikus RaRa-le, kes saab oma soovidele vastavalt parameetreid koostöös muuta. Kokkuvõtvalt on meie pakutud lahendus selgelt edukam hetkel parimatest kommertsmudelitest, kuid inimparandustele jääb alla.

Töö käigus on teostatud detailne andmestiku analüüs, mille käigus on kaardistatud probleemsed kohad, üldine kvaliteet ning manuaalselt autorite poolt parandatud testandmestik. Parandatud testandmestik koosneb 2001-st kontrollitud tekstipaarist, kus igale OCR tekstile vastab meie poolt kontrollitud ja parandatud tekst. Samuti on kombineerimise tulemusena saadud mitmeid treening- ja testandmestikke, mille abil saab teostada nii peenhäälestamist, eelistuste suunamist kui ka tekstide kvaliteedi hindamist. Valimi suurendamiseks ja mitmekesistamiseks on DIGAR-ist saadud tekstidele lisatud tekstipaare, kus OCR tekst on saadud sünteetiliste vigade genereerimisega.

Samuti on pakutud edasiarendustena tegevusi, mida antud töö käigus ei realiseeritud, näiteks alternatiivsete keelemudelite treenimine või parameetritega laialdasem katsetamine.

Lõputöös kasutusel olevad koodid, andmestikud ning tulemused on kättesaadaval GitHubis[3] ning kõik neli mudelit koos kasutamiseks vajaliku koodiga on kättesaadavad Hugging Face'is[4].

[3]GitHub-i repositoorium `https://github.com/mari-annam/estonian-ocr`
[4]Hugging Face `https://huggingface.co/mariannam`

Kasutatud kirjandus

[1] Eesti Rahvusraamatukogu. *OCR tekstiparandused*. [Loetud: 20-02-2025]. URL: https://digilab.rara.ee/andmestikud/ocr-tekstiparandused/.

[2] Eesti Rahvusraamatukogu. *DIGAR Eesti artiklid*. [Loetud: 27-03-2025]. URL: https://dea.digar.ee/?a=p&p=home&e=-------et-25--1--txt-txIN%7ctxTI%7ctxAU%7ctxTA----------.

[3] Eesti Rahvusraamatukogu. *Automaatselt tuvastatud teksti parandamine*. [Loetud: 27-03-2025]. URL: https://dea.digar.ee/?a=p&p=home&e=-------et-25--1--txt-txIN%7ctxTI%7ctxAU%7ctxTA----------#correcttext.

[4] Krister Kruusmaa. *Estonian Historical Newspaper Crowdsourced OCR Corrections [Data set]*. [Loetud: 15-05-2025]. 2024. URL: https://doi.org/10.5281/zenodo.13325713.

[5] Krister Kruusmaa, Laura Nemvalts, and Peeter Tinits. *Tehisintellekti põhised lahendused mäluasutuste töövoos*. [Loetud: 14-03-2025]. URL: https://docs.google.com/presentation/d/1ES2jvC0XIO2lWiqQswqlxesh3XfaE15l/edit#slide=id.p1.

[6] Sean Michael Kerner. *DeepSeek explained: Everything you need to know*. [Loetud: 27-03-2025]. URL: https://www.techtarget.com/whatis/feature/DeepSeek-explained-Everything-you-need-to-know.

[7] Inc. DeepSeek. *Models Pricing*. [Loetud: 27-03-2025]. URL: https://api-docs.deepseek.com/quick_start/pricing.

[8] Pankaj Tripathi. *The Brief History of OCR Technology*. [Loetud: 10-04-2025]. 2025. URL: https://www.docsumo.com/blog/optical-character-recognition-history.

[9] Cem Dilmegani. *State of OCR in 2025: Is it dead or a solved problem?* [Loetud: 14-03-2025]. URL: https://research.aimultiple.com/ocr-technology/.

[10] James Zhang et al. "Post-OCR Correction with OpenAI's GPT Models on Challenging English Prosody Texts". In: *Proceedings of the ACM Symposium on Document Engineering 2024*. 2024, pp. 1–4.

[11] Emanuela Boros et al. “Post-correction of historical text transcripts with large language models: An exploratory study”. In: *Proceedings of the 8th joint SIGHUM workshop on computational linguistics for cultural heritage, social sciences, humanities and literature (LaTeCH-CLfL 2024)*. Association for Computational Linguistics. 2024, pp. 133–159.

[12] Jenna Kanerva et al. “OCR Error Post-Correction with LLMs in Historical Documents: No Free Lunches”. In: *arXiv preprint arXiv:2502.01205* (2025).

[13] Callum Booth, Robert Shoemaker, and Robert Gaizauskas. “A Language Modelling Approach to Quality Assessment of OCR‘ed Historical Text”. In: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Ed. by Nicoletta Calzolari et al. Marseille, France: European Language Resources Association, June 2022, pp. 5859–5864. URL: https://aclanthology.org/2022.lrec-1.630/.

[14] Yves Maurer Pit Schneider. “Rerunning OCR: A Machine Learning Approach to Quality Assessment and Enhancement Prediction”. In: *Journal of Data Mining and Digital Humanities* (2022). ISSN: 2416-5999. URL: https://jdmdh.episciences.org/10239/pdf.

[15] Prashant Bansal. “Prompt Engineering Importance and Applicability with Generative AI”. In: *Journal of Computer and Communications*. 2024.

[16] Thi Tuyet Hai Nguyen et al. “Survey of Post-OCR Processing Approaches”. In: *ACM Comput. Surv.* 54.6 (July 2021). ISSN: 0360-0300. DOI: 10.1145/3453476. URL: https://doi.org/10.1145/3453476.

[17] Maxime Labonne. *Fine-tune Mistral-7b with Direct Preference Optimization*. [Loetud: 20-03-2025]. 2023. URL: https://mlabonne.github.io/blog/posts/Fine_tune_Mistral_7b_with_DPO.html.

[18] João Lages. *Direct Preference Optimization (DPO)*. [Loetud: 21-03-2025]. 2023. URL: https://medium.com/@joaolages/direct-preference-optimization-dpo-622fc1f18707.

[19] Aryaman Singh. *D.P.O vs R.L.H.F : A Battle for Fine-Tuning Supremacy in Language Models*. [Loetud: 21-03-2025]. 2023. URL: https://medium.com/@sinarya.114/d-p-o-vs-r-l-h-f-a-battle-for-fine-tuning-supremacy-in-language-models-04b273e7a173.

[20] Anukriti Ranjan. *Preference Tuning LLMs: PPO, DPO, GRPO — A Simple Guide*. [Loetud: 21-03-2025]. 2025. URL: https://anukriti-ranjan.medium.com/preference-tuning-llms-ppo-dpo-grpo-a-simple-guide-135765c87090.

[21] Hele-Andra Kuulmets et al. “Teaching Llama a New Language Through Cross-Lingual Knowledge Transfer”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Ed. by Kevin Duh, Helena Gomez, and Steven Bethard. [Loetud-16-03-2025]. Mexico City, Mexico: Association for Computational Linguistics, June 2024, pp. 3309–3325. DOI: 10.18653/v1/2024.findings-naacl.210. URL: https://aclanthology.org/2024.findings-naacl.210.

[22] Hasan Tanvir, Claudia Kittask, and Kairit Sirts. *EstBERT: A Pretrained Language-Specific BERT for Estonian*. [Loetud: 16-03-2025]. 2020. arXiv: 2011.04784 [cs.CL].

[23] Meta. *Llama-2-7b-hf*. [Loetud: 16-03-2025]. 2023. URL: https://huggingface.co/meta-llama/Llama-2-7b-hf.

[24] Hele-Andra Kuulmets et al. *Teaching Llama a New Language Through Cross-Lingual Knowledge Transfer*. [Loetud: 16-03-2025]. 2024. arXiv: 2404.04042 [cs.CL].

[25] Toloka. *Base LLM vs. nstruction-tuned LLM*. [Loetud: 16-03-2025]. 2024. URL: https://toloka.ai/blog/base-llm-vs-instruction-tuned-llm/.

[26] Agnes Luhtaru et al. “To Err Is Human, but Llamas Can Learn It Too”. In: *Findings of the Association for Computational Linguistics: EMNLP 2024*. Ed. by Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen. [Loetud: 16-03-2025]. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 12466–12481. DOI: 10.18653/v1/2024.findings-emnlp.727. URL: https://aclanthology.org/2024.findings-emnlp.727.

[27] Shuhao Guan et al. “Effective Synthetic Data and Test-Time Adaptation for OCR Correction”. In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*. 2024, pp. 15412–15425.

[28] *Pairwise sequence alignment*. [Loetud: 27-04-2025]. URL: https://biopython.org/docs/dev/Tutorial/chapter_pairwise.html.

[29] Saul B. Needleman and Christian D. Wunsch. “A general method applicable to the search for similarities in the amino acid sequence of two proteins”. In: *Journal of Molecular Biology* 48.3 (1970), pp. 443–453. ISSN: 0022-2836. DOI: https://doi.org/10.1016/0022-2836(70)90057-4. URL: https://www.sciencedirect.com/science/article/pii/0022283670900574.

[30] Kenneth Leung. *Evaluate OCR Output Quality with Character Error Rate (CER) and Word Error Rate (WER)*. [Loetud: 28-03-2025]. 2021. URL: https://towardsdatascience.com/evaluating-ocr-output-quality-with-character-error-rate-cer-and-word-error-rate-wer-853175297510/.

[31] Ethan Nam. *Understanding the Levenshtein Distance Equation for Beginners*. [Loetud: 21-03-2025]. URL: https://medium.com/@ethannam/understanding-the-levenshtein-distance-equation-for-beginners-c4285a5604f0.

[32] Nathan Lambert. *Reinforcement Learning from Human Feedback*. [Loetud: 02-04-2025]. Online, 2024. URL: https://rlhfbook.com.

[33] Haridus- ja Noorteamet. *How Is Python Used in Machine Learning?* [Loetud: 14-03-2025]. URL: https://builtin.com/machine-learning/python-machine-learning#:~:text=Python%20is%20the%20best%20choice,versatility%20to%20machine%20learning%20environments..

[34] Tallinna Tehnikaülikool. *Taltech AI-Lab*. [Loetud: 14-03-2025]. URL: https://ai-lab.pages.taltech.ee/en/intro/.

[35] Anthony Corbo. *Slurm ressursihaldur*. [Loetud: 20-03-2025]. URL: https://kuutorvaja.eenet.ee/wiki/Slurm_resursihaldur.

[36] Melanie Krause. *Google Colab: the power of the cloud for machine learning*. [Loetud: 20-03-2025]. URL: https://datascientest.com/en/google-colab-the-power-of-the-cloud-for-machine-learning#:~:text=Google%20Colab%2C%20short%20for%20Google,software%20installed%20on%20your%20computer..

[37] Python Software Foundation. *JiWER*. [Loetud: 21-03-2025]. URL: https://pypi.org/project/jiwer/.

[38] *pandas*. [Loetud: 21-03-2025]. URL: https://pandas.pydata.org/.

[39] *Matplotlib*. [Loetud: 21-03-2025]. URL: https://matplotlib.org/.

[40] *seaborn*. [Loetud: 21-03-2025]. URL: https://seaborn.pydata.org/.

[41] *Juhend:Korrektuur*. [Loetud: 28-03-2025]. 2015. URL: https://et.wikisource.org/wiki/Juhend:Korrektuur.

[42] Alan Thomas, Robert Gaizauskas, and Haiping Lu. "Leveraging LLMs for post-OCR correction of historical newspapers". In: *Proceedings of the Third Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA)@ LREC-COLING-2024*. 2024, pp. 116–121.

[43] Gemma Team. "Gemma 3". In: (2025). URL: https://goo.gle/Gemma3Report.

[44] Yuntao Bai et al. “Training a helpful and harmless assistant with reinforcement learning from human feedback”. In: *arXiv preprint arXiv:2204.05862* (2022).

Lisa 1 – Lihtlitsents lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks[1]

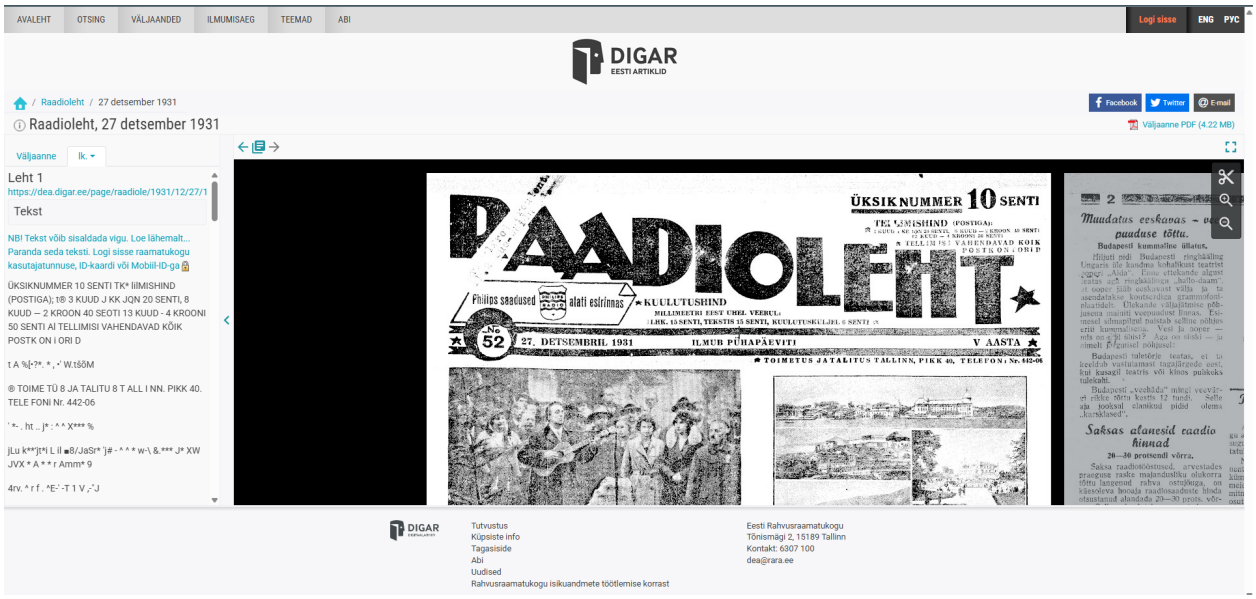
Mina, Mari-Anna Meimer ja mina, Loore Lehtmets

1. Anname Tallinna Tehnikaülikoolile tasuta loa (lihtlitsentsi) enda loodud teose "Ajalooliste eestikeelsete OCR tekstide järeltöötluse ja hindamise automatiseerimine Eesti Rahvusraamatukogu jaoks", mille juhendaja on Priit Järv
   1.1. reprodutseerimiseks lõputöö säilitamise ja elektroonse avaldamise eesmärgil, sh Tallinna Tehnikaülikooli raamatukogu digikogusse lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
   1.2. üldsusele kättesaadavaks tegemiseks Tallinna Tehnikaülikooli veebikeskkonna kaudu, sealhulgas Tallinna Tehnikaülikooli raamatukogu digikogu kaudu kuni autoriõiguse kehtivuse tähtaja lõppemiseni.
2. Oleme teadlikud, et käesoleva lihtlitsentsi punktis 1 nimetatud õigused jäävad alles ka autorile.
3. Kinnitame, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest ning muudest õigusaktidest tulenevaid õigusi.

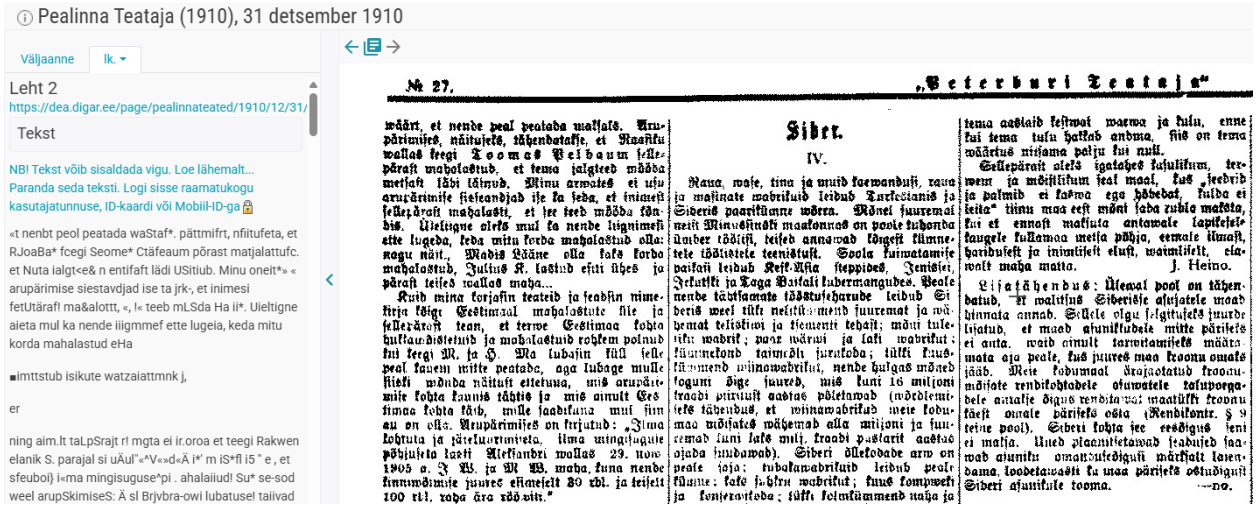
16.05.2025

[1] Lihtlitsents ei kehti juurdepääsupiirangu kehtivuse ajal vastavalt üliõpilase taotlusele lõputööle juurdepääsupiirangu kehtestamiseks, mis on allkirjastatud teaduskonna dekaani poolt, välja arvatud ülikooli õigus lõputööd reprodutseerida üksnes säilitamise eesmärgil. Kui lõputöö on loonud kaks või enam isikut oma ühise loomingulise tegevusega ning lõputöö kaas- või ühisautor(id) ei ole andnud lõputööd kaitsvale üliõpilasele kindlaksmääratud tähtajaks nõusolekut lõputöö reprodutseerimiseks ja avalikustamiseks vastavalt lihtlitsentsi punktidele 1.1. ja 1.2, siis lihtlitsents nimetatud tähtaja jooksul ei kehti.

Lisa 2 – Väljaanne digiarhiivi koduleheküljel



Joonis 12. *Kuvatõmmis väljaandest digiarhiivis.*



Joonis 13. *Kuvatõmmis halva kvaliteediga digiteerimisest digiarhiivis.*

Joonis 14. *Kuvatõmmis tekstituvastusest digiarhiivis.*

Joonis 15. *Kuvatõmmis kuulutest väljaandes.*

Lisa 3 – Viibad

```
prompt = f"""### Instruction:
Paranda OCR vead selles eestikeelses tekstis.

### Input:
{ocr}

### Response:
{gt}"""
```

Viip 3.1. Viip paranduste mudelile peenhäälestamisel.

```
prompt = f"Paranda OCR vead selles ajaloolises eestikeelses
    tekstis. Tekst on aastast {year}. Tagasta ainult parandatud
    tekst, ära kirjuta midagi muud, ära mõtle sõnu juurde. Kui sa
     teksti parandada ei oska, siis tagasta OCR tekst parandamata
     kujul. \nOCR TEKST: {ocr_text}"
```

Viip 3.2. Põhjalik viip ChatGPT-le parandamisel.

```
prompt = f"Paranda OCR vead selles ajaloolises eestikeelses
    tekstis. Tagasta ainult parandatud tekst. \nOCR TEKST: {
    ocr_text}"
```

Viip 3.3. Lühike viip ChatGPT-le parandamisel.

```
prompt = f"""### Instruction:
Kui suur protsent tähemärke sellest ajaloolisest eestikeelsest
    tekstist on vigane? Tagasta protsent täisarvuna.

### Input:
{ocr}

### Response:
{score_ocr}"""
```

Viip 3.4. Viip hindamismudelile, mis tagastab CER-i protsendina.

```
prompt = f"""### Instruction:
Kui suur on tõenäosus, et parandatud tekst on OCR tekstist parem
    ? Tagasta tõenäosus täisarvulise protsendina.

### Input:
OCR TEKST: {ocr}

PARANDATUD TEKST: {pred}

### Response:
{cer}"""
```

Viip 3.5. Viip hindamismudelile, mis tagastab tõenäosuse, et parandus on OCR tekstist parem.

```
prompt = f"""Kui suur protsent tähemärke sellest ajaloolisest
    eestikeelsest tekstist on vigane? Tagasta protsent täisarvuna
     vahemikus 0-100. Ära tagasta midagi muud. \nTEKST: {ocr_text
    }"""
```

Viip 3.6. Viip ChatGPT-le ja DeepSeek-ile CER-i ennustamisel

```
prompt = f"""Kui suur on tõenäosus, et parandatud tekst on OCR
    tekstist parem? Paranduse puhul on oluline, et ei lisataks sõ
    nu või lauseid. Tagasta tõenäosus täisarvulise protsendina
    vahemikus 0-100, ära tagasta midagi muud. 0 tähendab, et
    parandus on palju halvem OCR tekstist, 50 tähendab, et OCR
    tekst ja parandus on umbes võrdse kvaliteediga, 100 tähendab,
     et parandus on väga palju parem OCR tekstist. \nOCR TEKST: {
    ocr_text} \nPARANDATUD TEKST: {prediction}"""
```

Viip 3.7. Viip ChatGPT-le ja DeepSeek-ile paranduste hindamisel.

```
"prompt": [
    f"""### Instruction:
    Paranda OCR vead selles eestikeelses tekstis.

    ### Input:
    {ocr_text}

    ### Response:
```

```
    """
    for ocr_text in samples["ocr_text"]
],
"chosen": samples["chosen_response"],
"rejected": samples["rejected_response"]
```

Viip 3.8. Viip paranduste mudelile eelistuste suunamisel.

Lisa 4 – RaRa viip ChatGPT-4o mini mudelile

```
The following is the OCR output from a digitized historical
    Estonian newspaper from {year}. Analyze the text placed after
     "TEXT" and decide if it is reasonably free of OCR errors.
    Return a rating on the scale of 1 to 5.

5 - The text is clear and readable. It may contain unusual
    spellings and use of punctuation throughout, but there are no
     distorted words.
4 - The text is readable, but contains some distortions of
    alphabetical characters. These distortions do not impede
    understanding the text at any given point.
3 - The text is readable with minor difficulties. Words and
    phrases may be noticeably distorted.
2 - The text is only readable with great difficulties. All or
    almost all sentences contain severe errors that make it very
    hard to understand.
1 - The text is unreadable. It contains mostly gibberish and
    random symbols, almost no words are recognizable.

If you are hesitating between 4 and 5, it is probably a 5. If
    you are hesitating between 2 and 3, it is probably a 2.

Note: the use of "w" instead of "v" and "=" instead of "-" are
    elements of historical orthography an do not count as errors.

Do not reply anything else than a number from 1 to 5, unless
    explicitly asked to do so.

TEXT:
{ocr_transcription}"
```

Lisa 5 – Sünteetilised vead

Ground Truth

12. Isa tunnistus. 13. Nõusoleku-seletuse kättemuretsemine sõjawäkke wõtmiseks sõjawäeosa läbi.

OCR tekst

12. Ma tunniStuS. 13. Nõusoleku -seletuse kättemuretjemine jvjawäkke wõtmisekS sõjawäeosa läbi.

Sünteetilised vead (tõenäosuslik)

I2. isa tu»nistus. 13. Nõutollku seletufe kättemuretsemine sõj&wäkke wõtmrs-ks sõiamäeosa läbi.

Sünteetilised vead (juhuslik)

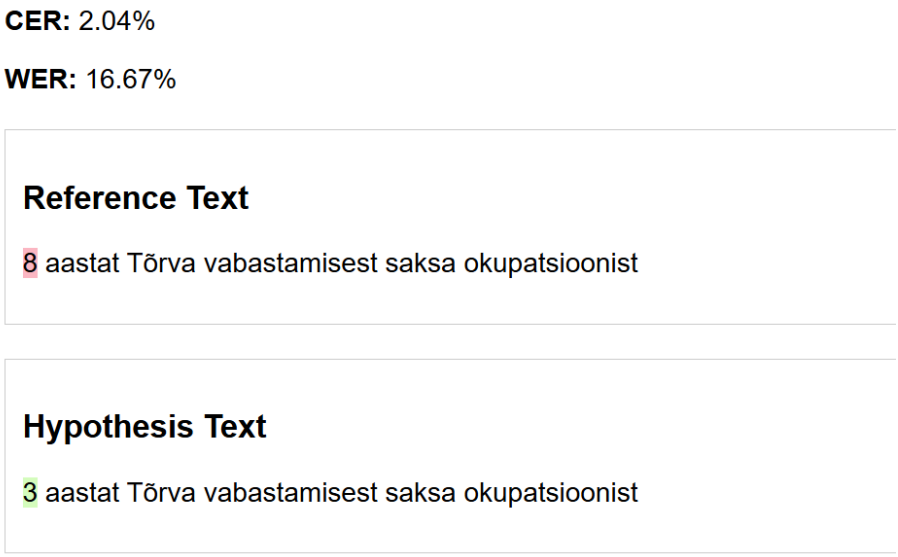
12. Isf tunnistus. 13. Nõusolekouäseletuse kättemuretsemine sõjawäkke wõtmiseks sõjaqäeos läbi.

Sünteetilised vead (enimlevinud)

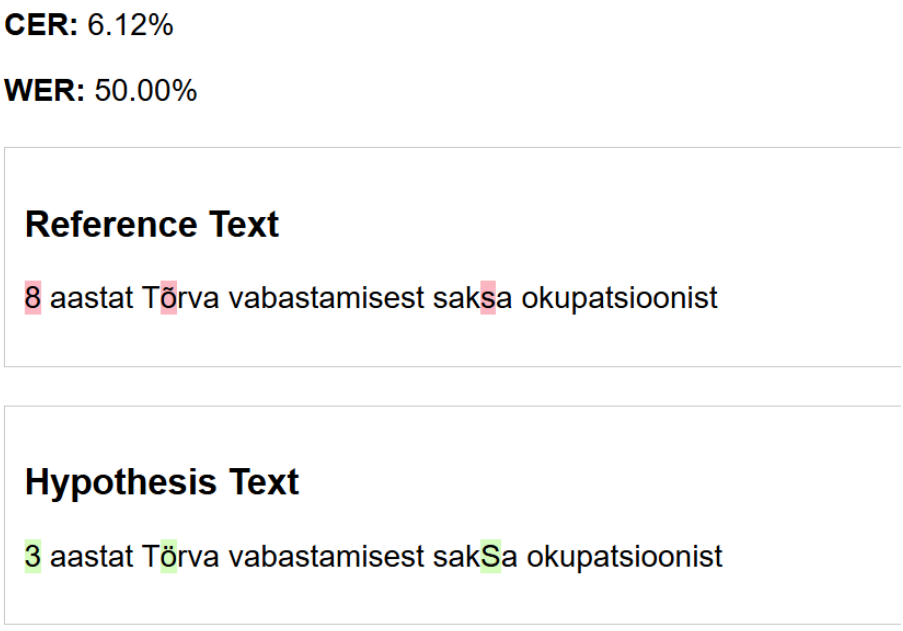
i2. Isa tunnistus. 13. Nõusoleku-seletuse kättemuretsem!ne sõjawäkke wõtmiseks sõjowäeosa läbi.

Joonis 16. *Erinevate vigade näide ühe teksti puhul.*

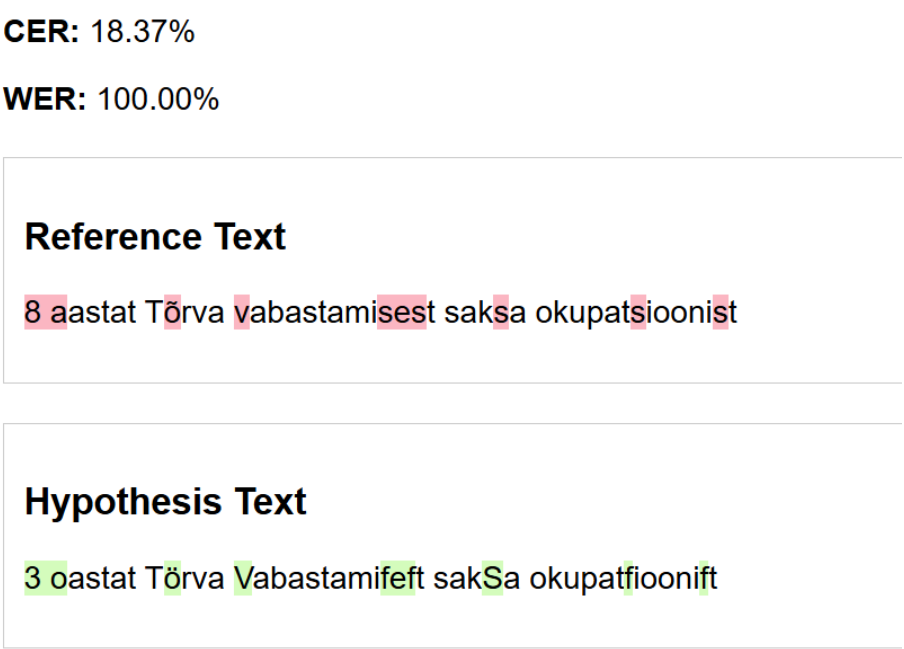
Lisa 6 – Müratasemed ühe lause näitel



Joonis 17. *Madal müratase.*



Joonis 18. *Keskmine müratase.*



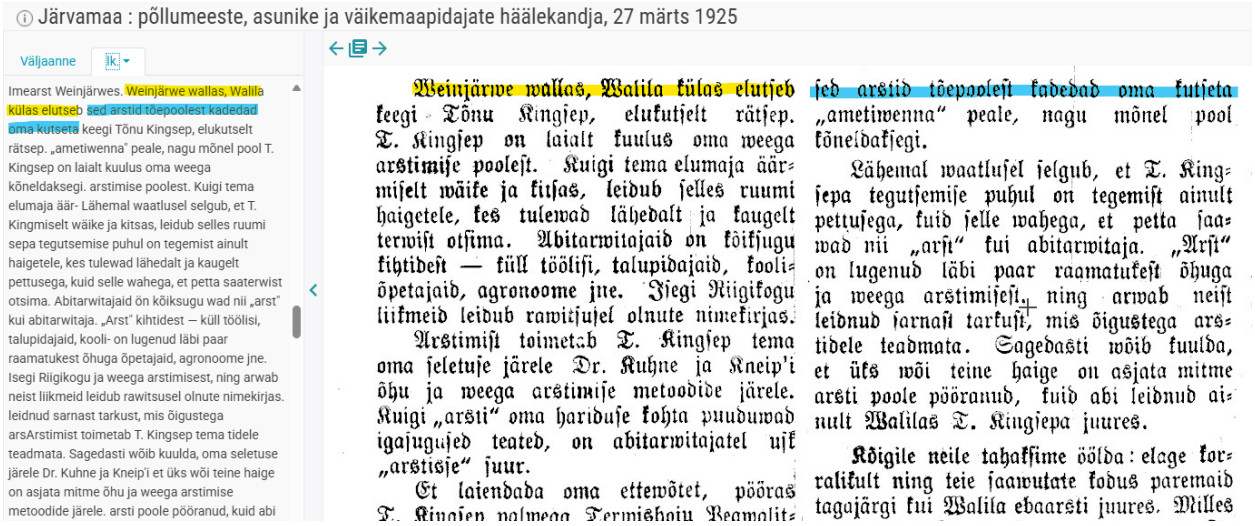
Joonis 19. *Kõrge müratase.*

Lisa 7 – Sünteetiliste vigade andmestik

Tabel 17. *Sünteetiliste vigade andmestik*

Teos	**Ilmumisaasta (eesti keeles)**	**Näidete arv**
Võrumaa jutud	1924-1933	3726
Valge ahv	1928	1824
Tõde ja õigus I	1926	3755
Minu sõbrad	1910	1206
Mahtra sõda	1902	4150
Kuritöö ja karistus	1929	4265
Kõrboja peremees	1922	2145
Dorian Gray portree	1957	5863
Eesti Päevaleht	1991	555

Lisa 8 – Segunenud veerud



Joonis 20. *Segunenud veerud liiga lähestikku paiknevate veergude tõttu.*