TARTU ÜLIKOOL
HUMANITAARTEADUSTE JA KUNSTIDE VALDKOND
EESTI JA ÜLDKEELETEADUSE INSTITUUT

Marko Sasi

KEELETUVASTAJATE VÕRDLUS

Bakalaureusetöö

Kadri Muischnek, PhD

TARTU 2026

Autorsuse kinnitus

Kinnitan, et olen käesoleva lõputöö ise kirjutanud ning toonud korrektselt välja teiste autorite panuse. Töö on kirjutatud, lähtudes Tartu Ülikooli humanitaarteaduste ja kunstide valdkonna lõputööde nõuetest, ning on kooskõlas heade akadeemiliste tavadega, sealhulgas on järgitud teaduseetika põhimõtteid.

Marko Sasi

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärk oli võrrelda kahteteist laialdaselt kasutusel olevat keeletuvastajat ning hinnata nende sobivust eesti-inglise koodivahetuse sõnatasandi tuvastamiseks. Töös testiti tuvastajaid AnE-LID, lingua (sealhulgas režiim lingua-mixed), heliport, GlotLID, nb-nordic-lid, fastText, fast-langdetect, pycld2, masklid (sealhulgas režiim masklid-cs) ja langid. Hindamiseks koguti, töödeldi ja märgendati käsitsi eesti-inglise koodivahetusega lausete korpus Redditi alafoorumist r/Eesti[1] (534 lauset, 9 773 sõnet, neist 20,41% ingliskeelsed). Võrdluskorpusena kasutati projekti „Teismeliste Keel Eestis" raames kogutud Tšätikorpust (61 333 sõnet, neist 7,26% ingliskeelseid). Tuvastajaid hinnati sõne õigsuse, täpsuse, saagise, F1-skoori ja koondskoori põhjal. Parima koondtulemuse saavutas Tšätikorpusel heliport (74,32%) ja Redditi korpusel AnE-LID (93,49%); AnE-LID jäi Tšätikorpusel teiseks (71,49%) ning lingua-mixed Redditi korpusel teiseks (87,63%) ja Tšätikorpusel neljandaks (64,53%). Konservatiivsed tuvastajad (heliport, GlotLID, nb-nordic-lid) saavutasid kõrge täpsuse, kuid madala saagise; agressiivsed tuvastajad (masklid, masklid-cs, langid) vastupidi. Töös pakutakse välja soovitused erinevateks kasutuskontekstideks.

Võtmesõnad: koodivahetus, keeletuvastus, korpuslingvistika, sotsiaalmeedia

[1] https://www.reddit.com/r/Eesti/

Sisukord

Sissejuhatus

Eesti-inglise koodivahetus, ehk olukord, kus eestikeelses tekstis esineb ingliskeelseid sõnu või väljendeid, on viimastel aastatel suurt kõneainet tekitanud. Enim levib selline nähtus just noorte omavahelises suulises ning internetisuhtluses. Käesolevas töös mõistetakse koodivahetuse all olukorda, kus ühe lause piires esinevad kahest erinevast keelest pärinevat keeleüksused (Das & Gambäck 2013). Selles töös kasutatud materjalis esineb ainult eesti-inglise koodivahetust.

Keeletuvastaja on tarkvara, mis määrab kindlaks, millises keeles on kirjutatud dokument või selle osa. Käesoleva bakalaureusetöö eesmärk on võrrelda kahteteist laialdaselt kasutusel olevat keeletuvastajat – ane-id, lingua, heliport, GlotLID, nb-nordic-lid, fastText, fast-langdetect, pycld2, masklid ja langid. Lisaks eelnevalt välja toodud kümnele keeletuvastajale kasutati töös lingua ja masklidi spetsiifilisi koodivahetuse tuvastamise režiime – lingua-mixed ning masklid-cs, mistõttu töös võrreldakse kokku kahtteist keeletuvastajat. Eesmärk on hinnata keeletuvastajate sobivust eesti-inglise koodivahetuse sõnatasandi tuvastamiseks, st millise kvaliteediga suudavad need tarkvarad määrata kindlaks, milline sõna koodivahetusega lauses on eesti ja milline inglise keeles.

Eesmärgi saavutamiseks koguti, töödeldi ning märgendati eesti-inglise koodivahetusega lausete korpus Redditi alafoorumist r/Eesti, mida võrreldakse hindamisel Virve-Anneli Vihmani jt poolt projekti "Teismeliste keel Eestis" (TeKE[2]) raames koostatud Eesti teismeliste keele korpusega. Töö hõlmab kolme omavahel seotud eesmärki:

(1) võrrelda keeletuvastajate tulemusi kahel iseloomult erineval korpusel ning pakkuda võrdluse põhjal soovitusi erinevateks kasutuskontekstideks;

(2) luua Redditi tekstidest piisavalt mahukas ja mitmekesine korpus analüüsiks;

(3) valmistada Redditist kogutud materjal ette käsitsi märgendamiseks.

Tuvastajate hindamiseks kasutatakse kahte korpust. Esimene on Vihmani jt poolt kogutud Eesti teismeliste keele korpus, mis sisaldab peamiselt Facebook Messengeri eravestluseid noorte eesti keele kõnelejate seas. Eesti teismeliste keele korpuse tekstid on lühikesed, mitteametlikud ning sisaldavad rohkesti lühendeid ja kõnekeelseid vorme. Teiseks korpuseks on Redditi alafoorumist r/Eesti kogutud ning käsitsi märgendatud laused, mis on avalikud ja keskmiselt

[2] https://sisu.ut.ee/teke/

pikemad ning ortograafiliselt korrektsemad kui sõnumivestlused. Korpuste märgendamisel järgitakse Eesti teismeliste keele korpuse märgendusskeemi, mida on Redditi eripärasid arvestades vähesel määral kohandatud. Tuvastajate kvaliteeti hinnatakse sõnetasandi täpsuse, saagise ja F1-skoori ning nende põhjal arvutatud koondskoori alusel.

Töö algab ajaloolise ja teoreetilise ülevaatega keeletuvastusest ning töös võrreldavate keeletuvastajate kirjeldustest, kus tutvustatakse iga keeletuvastaja tunnuseid, masinõppe lähenemist ning treening- ja testkorpust. Sellele järgneb kasutatud korpuste peatükk, milles kirjeldatakse Redditi andmete kogumise arhitektuuri, töövoo skripte, eetilisi kaalutlusi ning mõlema korpuse peamisi numbrilisi näitajaid. Sellele järgneb Redditi korpuse käsitsi märgendamise põhimõtete kirjeldus, mis tugineb Eesti teismeliste keele korpuse märgendusele. Seejärel kirjeldatakse lühidalt teismeliste keele korpuse allosa - Tšätikorpust ning võrreldakse seda eelnevalt koostatud Redditi korpusega. Töö viimases osas esitatakse tulemused – tuuakse välja keeletuvastajate koondtabelid mõlema korpuse kohta, võrdlev analüüs ning soovitused erinevateks kasutuskontekstideks.

1. Keeletuvastus ja keeletuvastajad

1.1. Lühike ülevaade keeletuvastuse ajaloost

Järgnev põgus ülevaade keeletuvastuse ajaloost põhineb valdavalt Tommi Jauhiaineni, Marco Lui, Marcos Zampieri, Timothy Baldwini ja Krister Lindéni (2019) ülevaateartiklil „Automatic Language Identification in Texts: A Survey", üksikute meetodite ja konkreetsete tulemuste täpsustamiseks on viidatud ka algallikatele. Esimesed katsetused tekstipõhises keeletuvastuses pärinevad aastast 1960 (Mustonen 1965), kuid aktiivsem valdkonna areng algas 1980. aastatel, mil keskenduti teksti keelelise kuuluvuse automaatsele määramisele peamiselt tähestiku ja sagedaste sõnavormide alusel (Jauhiainen jt 2019). Algsed katsetused tuginesid käsitsi koostatud sõnaraamatutele ning tuvastasid keele sagedaste sõnade kattuvuse kaudu. Sellised lähenemised osutusid piiratuks, kuna need ei tulnud toime ortograafilise varieeruvuse ja lühikeste tekstidega.

Muutus toimus 1990. aastate alguses, kui hakati kasutama keelt iseloomustavate tähejärjendite (n-grammide) esinemissagedusi. Ted Dunning (1994) näitas, et Markovi ahelatel, kus iga tähe tõenäosus sõltub eelnevatest tähtedest, põhinev keeltevõrdlus võimaldab usaldusväärselt tuvastada keelt juba suhteliselt lühikestest tekstilõikudest. Valdkonna arengule aitas kaasa William B. Cavnari ja John M. Trenkle'i (1994) n-grammipõhine meetod, mis põhines tähe-n-grammide sagedusel. Iga keele kohta koostati treeningkorpuse põhjal kõige sagedasemate n-grammide nimekiri ning sissetulev tekst omistati sellele keelele, millise n-grammide profiiliga see kõige paremini ühildus. Meetodi lihtsus, ülekantavus keelte vahel ja sõltumatus morfoloogilisest eeltöötlusest tegid sellest pikaks ajaks valdkonna kuldstandardi. Gregory Grefenstette (1995) pakkus välja ka alternatiivse lähenemise, mis tugines lühikeste sõnade sagedustele, ning näitas, et trigrammipõhine ja lühikeste sõnade põhine meetod annavad pikematel tekstidel võrreldavaid tulemusi, samas kui väga lühikestel tekstidel saavutasid n-grammid kindlama täpsuse.

2000. aastatel liikus keeletuvastamine eelmainitud meetoditest masinõppepõhiseks, mis rakendas üldotstarbelisi klassifikaatoreid – naiivset Bayesi klassifikaatorit, tugivektormasinaid ja logistilist regressiooni. Sellised meetodid võimaldasid ühendada tunnuseid, nagu sõna- ja tähe-n-gramme, sõnapikkusi ning ortograafilisi mustreid. Marco Lui & Timothy Baldwin (2012) demonstreerisid, et ülesanne läheb raskemaks keelte arvu suurenedes, treeningandmete mahu

vähenedes ja dokumentide pikkuse lühenedes. Lui & Baldwini poolt välja töötatud rakendus *langid.py* koondas need arengud avatud lähtekoodiga rakenduseks, mis oli treenitud 97 keelel ning on senini üks laialdasemalt kasutatud keeletuvastajaid. Selle perioodi keskseks arusaamaks sai, et keeletuvastuse täpsus sõltub tugevalt treeningandmetest ja teksti pikkusest. Pikkade ja normaliseeritud tekstide puhul saavutasid parimad tööriistad üle 99-protsendilise täpsuse, ent lühikestel sotsiaalmeediatekstidel ja lähisugulaskeelte paaridel langes täpsus märgatavalt (Lui & Baldwin 2014).

Sügavõppe esiletõus 2010. aastate teisel poolel mõjutas ka keeletuvastust, ehkki n-grammipõhised süsteemid olid juba jõudnud kõrge täpsuseni. Armand Joulini jt (2017) välja töötatud *fastText* tuli välja kompaktse närvivõrgupõhise liigitajaga, mis kasutas n-grammidel põhinevaid sõnavektoreid ning võimaldas suurel kiirusel tuvastada keelt. Paralleelselt katsetati tähemärgitasandi konvolutsioonilisi ja rekurrentseid närvivõrke, mille hulgas Aaron Jaech jt (2016) esitasid hierarhilise märgi-sõna mudeli, mis arvestas paremini morfoloogiliselt rikaste keelte ja mittestandardsete ortograafiaga sotsiaalmeediatekstides. Tommi Jauhiaineni jt (2019) põhjalik ülevaade näitas, et kuigi sügavõppepõhised süsteemid saavutavad piisavalt suurte andmestikega parima tulemuse, jääb lühikeste tekstide ja lähisugulaskeelte eristamine endiselt problemaatiliseks.

Keeletuvastajate rakendamine koodivahetuse kontekstis tõi esile uue väljakutse, sest traditsiooniliselt eeldati üht keelt kogu sisendtekstis, samal ajal kui koodivahetus nõuab sõna- või fraasitasandi analüüsi. Dong Nguyen & A. Seza Doğruöz (2013) olid esimeste seas, kes uurisid sõnatasandi keeletuvastust veebivestluste tingimustes, näidates, et dokumenditasandi tööriistad ei ole sellise ülesande jaoks piisavad. On näidatud, et isegi tipptasemel sõnatasandi keeletuvastajad kaotavad oluliselt täpsust lühikeste ja ortograafiliselt mittestandardsete sotsiaalmeediatekstide puhul, kus pärisnimed, laenud ja kirjavead on tihti raskesti eristatavad tegelikust koodivahetusest (Barman jt 2014). Eesti keele kontekstis on koodivahetuse automaatne tuvastamine jäänud seni pigem tahaplaanile, kuigi eesti-inglise keelekontakti avaldumisvorme on uuritud üldisemalt (Verschik & Kask 2019).

1.2. Keeletuvastajate kirjeldused

Käesolevas osas antakse ülevaade töös kasutatud keeletuvastajate omadustest, metoodilistest lähenemistest, treening- ja testkorpustest ning väljundist. Keeletuvastajad erinevad üksteisest muuhulgas selle poolest, kas nad määravad sisendteksti keele teksti kaupa (kogu sisend saab ühe märgistuse), lause kaupa (igale lausele eraldi märgistus) või sõne kaupa (iga sõne saab oma märgistuse). Iga allpool kirjeldatud keeletuvastaja juures on püütud välja tuua, milliseid tunnuseid ja masinõppe meetodeid mudel kasutab, kas autorid on avaldanud infot keeletuvastaja kvaliteedinäitajate ja treening- ning testkorpuse kohta ning kuidas keel väljundis märgendatakse. Paljude keeletuvastajate puhul puudub sellekohane avalik info ning need on allpool eraldi välja toodud.

1.2.1. AnE-LID

AnE-LID[3] (Any-English Code-Switching Language Identification) on sõnaklassifikaatori põhimõttel töötav keeletuvastaja, mis tuvastab inglise keelega seotud koodivahetust mitmekeelses tekstis (Sterner 2024). XLM-RoBERTa-large[4] keelemudeli peal on peenhäälestatud (*finetuned*) sõnaklassifikaatorikiht, mis omistab igale sõnele ühe neljast märgendist: *English*, *notEnglish*, *Mixed* (sõna, mis sisaldab kahe keele morfeeme, nt saksa-inglise *rewatchen*) ja *Other* (kirjavahemärgid jms). Masinõppe tüüp on seega juhendatud närvivõrgul põhinev tokenklassifitseerimine. Treeningandmed on koondatud mitmest olemasolevast koodivahetuse korpusest, mille märgendus on teisendatud ühtsesse märgendusskeemi ning mille koondamisprotsessi Sterner oma artiklis kirjeldab. Testimine toimus Igor Sterneri poolt koondatud andmestikul, mis sisaldab ka treeningus mitte kasutatud keelepaare. Tuvastamine toimub sõnekaupa. Väljundiks on iga sõne kohta eelnimetatud nelja klassi märgend koos vastava softmax-tõenäosusega, mis näitab mudeli kindlust selle klassi suhtes.

[3] https://huggingface.co/igorsterner/AnE-LID
[4] https://huggingface.co/FacebookAI/xlm-roberta-large

1.2.2. Lingua

Lingua[5] (Stahl 2024) on tõenäosuslikul n-grammi mudelil põhinev keeletuvastaja. Tuvastamine käib kahes etapis: esmalt otsitakse sisendist keelele tüüpilisi unikaalseid tähemärke, et välistada võimatud keeled, ning seejärel rakendatakse statistilist Naive Bayes tüüpi mudelit tähemärgi n-grammidel pikkustega 1–5. Masinõppe tüüp on seega juhendatud statistiline klassifitseerimine, närvivõrke kasutamata. Nii treening- kui testkorpus on autori poolt dokumenteeritud: andmed tulevad Wortschatzi[6] korpustest, mis on koostatud erinevate keelte uudisteportaalide põhjal (igas korpuses ligikaudu miljon lauset). Testimiseks kasutati eraldi dokumente erinevatelt veebilehekülgedelt (igas keeles u 10 000 lauset), millest omakorda valiti juhuslikult 1000 sõna, 1000 sõnapaari ja 1000 lauset ühtseks kvaliteedihinnanguks. Autor toob välja kvaliteedinäitajad keele kaupa (täpsus ehk *accuracy* sõnadel, sõnapaaridel ja lausetel eraldi ning nende keskmine, mediaan ja standardhälve) ning võrdleb Linguat Langdetecti, Langidi, Simplemma, CLD2 ja CLD3-ga. Näiteks saksa keeles on Lingua keskmine õigsus 89,3%. Tuvastamine toimub vaikimisi tekstikaupa. Märgendus on keelepõhine ning API ja mudelite kataloogistruktuur tuginevad ISO-koodidele (ISO 639-1 / ISO 639-3).

1.2.3. lingua-mixed

lingua-mixed ei ole eraldiseisev mudel, vaid Peter M.Stahli Lingua-teegi mitmekeelse teksti tuvastamise režiim (meetod *detect_multiple_languages_of*). Aluseks on sama statistiline n-grammi mudel, mida kasutab tavaline Lingua. Eraldi treeningut või treeningandmeid selle režiimi jaoks ei ole - kasutatakse Lingua üldisi keelemudeleid, mis on treenitud Wortschatzi korpustel. Funktsiooni peale on ehitatud segmenteerimisloogika, mis proovib tuvastada keelte vaheldumist ühes tekstis. Segmendi piirid määratakse konkureerivate keeleskooride põhjal. Autor ei too välja selle spetsiifilise režiimi kohta eraldi kvaliteedinäitajaid ning rõhutab dokumentatsioonis, et tegemist on eksperimentaalse funktsiooniga.

[5] https://github.com/pemistahl/lingua-py
[6] https://www.wortschatz.uni-leipzig.de/en

1.2.4. heliport

Heliport[7] on HeLI-OTS 2.0[8] *Rust-port*, mida saab kasutada Pythoni teegina (Zaragoza-Bernabeu 2024). Tunnusteks on terved sõnad ja tähemärgi n-grammid pikkustega 1–6 ning tuvastus toimub generatiivse tagasilangemise (*backoff*) põhimõttel – esmalt antakse sõnale hinnang sõnamudelist, selle puudumisel taganetakse 6-grammidele ja seejärel lühematele n-grammidele, puuduva tunnuse korral rakendub fikseeritud karistusskoor 7,0. Masinõppe tüüp on seega statistiline, relatiivsetel esinemissagedustel põhinev generatiivne mudel (närvivõrke ei kasutata). Mudel kasutab samu keelemudeleid mis HeLI-OTS 2.0 Java-versioon, mille on välja töötanud Tommi Jauhiainen, Heidi Jauhiainen ja Krister Lindén (Jauhiainen jt 2022). Keelemudelid on koostatud 200 keele jaoks ning sisaldavad iga keele kohta nii sõnasagedusi (terveid sõnu) kui ka tähemärgi n-gramme pikkustega 1–6. Treeningandmed pärinevad valdavalt Leipzigi korpustekogust (*Leipzig Corpora Collection*). Korpused on väga erineva suurusega ning osa neist on käsitsi puhastatud. Eraldi kvaliteedinäitajaid heliport-i autor välja ei too, kuid väidab, et tulemused on HeLI-OTS 2.0 Java-versiooniga "peaaegu identsed". Java-versiooni kohta raporteerivad Jauhiainen jt (2022) järgmised tulemused: ULI 110 testkorpusel (1 074 939 lauset 110 keeles, mis on ühised fastTextiga) saavutab HeLI-OTS makro-F1 skoori 0,699 ja mikro-F1 skoori 0,993. Kiiruse poolest on heliport originaalist ligikaudu 25 korda kiirem. Eeltöötluses esineb mõningaid erinevusi originaaliga (nt heliport ei eemalda URL-e ega @-märgiseid), mis võib anda üksikuid erinevaid tulemusi. Väljund on ISO 639-3 kood ühe keele kohta, valikuliselt koos skooriga. Tuvastus toimub teksti kaupa.

1.2.5. GlotLID

GlotLID[9] on fastText[10]-il põhinev keeletuvastusmudel, mis on spetsiaalselt välja töötatud vähese ressurssidega keelte katvuse parandamiseks (Kargaran jt 2023). Tunnused on tähemärgi n-grammid ja sõna alamosade (*subword*) n-grammid, mille peal töötab lineaarne multinoomne

[7] https://github.com/ZJaume/heliport
[8] https://zenodo.org/records/10907468
[9] https://github.com/cisnlp/GlotLID
[10] https://github.com/facebookresearch/fastText

logistiline regressioon (masinõppe tüüp on seega juhendatud klassifitseerimine fastText-klassifikaatoriga). Treeningkorpus GlotLID-C on autorite poolt üksikasjalikult kirjeldatud. See sisaldab umbes 289 miljonit lauset (kokku 40 GB) ja katab 1832 keelt. Allikad on valdavalt Vikipeedia artiklid, religioossed tekstid, tõlkekogumikud, jutustuste kogud ja uudisteportaalid. Veebiallikaid välditi teadlikult. Andmetest 85% kasutatakse treenimiseks ja ülejäänud 15% pealt võetakse igale keelele testkorpuseks kuni min(1000, n_l) lauset. Amir Hossein Kargaran jt (2023) toovad kvaliteedinäitajad välja järgmiselt: GlotLID-M ületab UDHR-i[11] põhjal koostatud testkorpusel varasemaid baaslahendusi (CLD3, FT176, OpenLID, NLLB) rohkem kui 12% absoluutse F1 poolest, säilitades samaaegselt madala valetuvastuse (*false-positives*) määra (FPR 0,0002 GlotLID-C-l ja 0,0010 FLORES-el). Tuvastus toimub tekstikaupa, väljundid järgivad ISO 639-3 formaati.

1.2.6. nb-nordic-lid

Nb-nordic-lid[12] on Norra rahvusraamatukogu tehisintellekti laboratooriumi (NbAiLab) treenitud fastText-tüüpi mudel, mis keskendub põhjamaade keeltele, sealhulgas saami keeltele (NbAiLab 2023). Tunnused on tähemärgi n-grammid ja sõna alamosa (*subword*) n-grammid ning masinõppe tüüp on juhendatud lineaarne klassifitseerimine fastText baasil. Mudel on saadaval kahes variandis: nb-nordic-lid tuvastab 12 enim levinud Põhja-Euroopa keelt (pluss inglise keel), nb-nordic-lid.159 laiendab katvust 159 keeleni. Treeningkorpust on vaid üldsõnaliselt kirjeldatud - kasutati lauseid GiellaTekno[13] tõlkemälust (ingl *Translation memories*) ja Wortschatzi korpustest, kuid täpsed laused, mahud ja allikate jaotused keelte kaupa ei ole avalikult kättesaadavad. Eraldi testkorpust ega süstemaatilist kvaliteedinäitajate tulemusi autorid ei avalda. Hugging Face'is on näidatud vaid üksikuid näidiseid. Tuvastamine käib tekstikaupa ja väljund on fastTextile omane *__label__xx* märgend, kus keelekoodid järgivad ISO 639-3 standardit.

1.2.7. fastText

FastText on Facebook AI Research'i poolt välja töötatud teek, mille keeletuvastusmudel *lid.176* tuvastab 176 keelt (Joulin jt 2017). Tunnusteks on tähemärgi n-grammid ja sõna alamosa

[11] UDHR – ÜRO inimõiguste ülddeklaratsioon, mille tõlked enam kui 500 keeles on keeletuvastuses laialt levinud testandmestik (https://www.unicode.org/udhr/)

[12] https://huggingface.co/NbAiLab/nb-nordic-lid

[13] https://giellatekno.uit.no/index.eng.html

(*subword*) tunnused, mille peal töötab lineaarne klassifikaator. Otsus tehakse kaalutud summa põhjal (skoor = $\text{tunnus}_1 \times \text{kaal}_1 + \text{tunnus}_2 \times \text{kaal}_2 + ...$, võitjaks saab suurima skooriga klass). Masinõppe tüüp on juhendatud mitmeklassiline logistiline regressioon hierarhilise softmaxiga, mis on koolitatud gradientlaskumise meetodil. Treeningkorpus on dokumenteeritud järgnevalt: andmed pärinevad Vikipeediast, Tatoebast[14] (erinevate keelte lausete ja tõlgete korpus) ja SETimes'ist (Kagu-Euroopa uudisteportaal, mis suleti 2015. aastal). Testkorpust ega eraldi kvaliteedinäitajaid ametlikus dokumentatsioonis välja toodud ei ole. Kvaliteet on raporteeritud vaid üldisel kujul Joulini jt (2017) artiklis muude tekstiklassifitseerimisülesannete raames, mitte keelte kaupa. Tuvastamine toimub tekstikaupa, väljundisildid järgivad ISO 639-1 koodide süsteemi ja esitatakse kujul *__label__xx* koos tõenäosusskooriga.

1.2.8. fast-langdetect

Fast-langdetect[15] on Pythoni teek, mis pakub kiireid ja mugavaid liideseid Facebooki fastTexti *lid.176* mudelile. Tegemist ei ole omaette arendusega, vaid optimeeritud sama fastTexti mudeli ümber, mida on kirjeldatud fastTexti alapeatükis. Tunnused, algoritm ja masinõppe tüüp on seega samad mis fastTextil. Teegi autorid eraldi treening- ega testkorpust välja ei too, need on samad fastTexti omadega. Kvaliteedinäitajatena viidatakse README-s üldisele fastTexti täpsusele „kuni 95%" ja teegi enda puhul umbes 80-kordsele kiirendusele võrreldes fastTexti teegiga, kuid kvaliteedihindamist oma andmestikul ei esitata. Tuvastamine toimub tekstikaupa ning väljundisildid on BCP-47 stiilis koodid (nt *en*, *zh-cn*, *pt-br*).

1.2.9. pycld2 (CLD2)

Pycld2[16] on Pythoni liides Google'i CLD2[17] (Compact Language Detector 2) teegile (McCandless 2013). Tunnusteks on tähemärgi n-grammid ja osaliselt terved sõnad, nende pealt töötab Naive Bayes klassifikaator. Masinõppe tüüp on seega statistiline klassifitseerimine koos heuristilise eeltöötlusega. Autor ei ole treeningkorpust dokumenteerinud üksikasjalikult – märgitud on vaid, et algset käsitsi valitud veebilehtede kogumit täiendati hilisemate hoolikalt kogutud veebikogumitega (umbes 100 miljonit veebilehte), et katta rohkem keeli. Eraldi

[14] https://tatoeba.org/en/
[15] https://github.com/LlmKira/fast-langdetect
[16] https://github.com/aboSamoor/pycld2
[17] https://github.com/CLD2Owners/cld2

testkorpust ega kvaliteedinäitajaid CLD2 autorid ise välja ei too. Tähelepanuväärne on, et CLD2 suudab mitmekeelsete tekstide korral tagastada ühe sisendi kohta kuni kolm keelt koos osakaaludega, põhiline kasutustase on aga teksti kaupa. Väljundiks on ISO 639-1 keelekoodid koos protsendiga

1.2.10. MaskLID

MaskLID-i[18] idee on järgmine: kui lauses on segamini kaks keelt L1 ja L2, siis tavaline softmax-klassifikaator tagastab peaaegu alati vaid domineeriva keele L1. MaskLID maskeerib L1-ga seotud tähemärgi n-grammid ning klassifitseerib teksti uuesti. Nii on tõenäolisem, et teisel ringil ennustatakse L2. Masinõppe tüüp ja treeningkorpus pärinevad GlotLID-ist. Meetod ise ei vaja lisatreeningut, seega eraldi treeningkorpust ei ole. Kargaran jt (2024) raporteerivad kvaliteedinäitajad sama testkorpuse peal kui masklid-cs (täpsed vasted, osalised vasted, valetuvastused) Väljundid järgivad aluseks oleva GlotLID-i formaati: ISO 639-3 kood koos skripti märgendiga (nt *__label__eng_Latn*).

1.2.11. MaskLID-cs

Masklid-cs on MaskLID koodivahetuse (*code-switching*) tuvastamise režiim, milles on režiim *predict_codeswitch* (Kargaran jt 2024). Tavalise MaskLID-iga võrreldes on eesmärk saada vastuseks mitu keelesilti ühe sisendi kohta koos vastavate tekstiosadega. Eraldi treeningkorpust sellel režiimil ei ole, kehtivad alusmudeli treeningandmed. Testkorpusena ja kvaliteedihinnangu alusena kasutavad autorid nelja koodivahetuse andmestikku (türgi-inglise, hindi-inglise, nepali-inglise, baski-hispaania) ning toovad välja kvaliteedinäitajad täpsete vastete (*exact match*) ja osaliste vastete (*partial match*) arvuna ning valetuvastuste arvuna. Näiteks türgi-inglise koodivahetuses tuvastas MaskLID-cs parimal juhul 91 koodivahetust 100-st, samas kui MaskLID-cs-ta alusmudel tuvastas neist vaid 4.

1.2.12. langid (langid.py)

Langid[19] (täpsemalt langid.py) on Marco Lui ja Timothy Baldwini (2012) loodud keeletuvastaja, mis on eeltreenitud 97 keele jaoks. Tunnuseks on tähemärgi n-grammide registrist sõltumatu

[18] https://github.com/cisnlp/masklid
[19] https://github.com/saffsd/langid.py

valik, mille tegemiseks kasutatakse autorite varasemas töös väljatöötatud tunnuste valikumeetodit. Masinõppe tüüp on multinoomne Naive Bayes klassifikaator. Treeningandmed on täpselt dokumenteeritud: andmed on kogutud viiest eri allikast – JRC-Acquis (ELi õigusaktide paralleelkorpus), ClueWeb09, Vikipeedia, Reuters RCV2 ning Debiani tarkvara lokaliseerimise korpus (Debian i18n). Lui & Baldwin (2012) raporteerivad kvaliteedinäitajad järgnevalt: langid.py on võrreldud viie pikemate dokumentide andmestiku ja kahe mikroblogi-andmestiku peal, raporteeritud on täpsus (*accuracy*) ja F1-skoor. Tulemused näitasid, et langid.py säilitab ühtlaselt kõrge täpsuse kõikide allikate peal. Tuvastamine toimub tekstikaupa, väljundi moodustavad keelekood ja tõenäosushinnang.

2. Kasutatud korpused

Käesolevas töös kasutatakse kahte koodivahetusega korpust – Redditi korpust ja Tšätikorpus. Redditi korpuse on kogunud töö autor ise, mistõttu peatükk algab selle korpuse kogumismetoodika kirjeldusega (alapeatükid 2.1–2.7). Seejärel tutvustatakse mõlemat korpust ja esitatakse nende võrdlus (alapeatükk 2.8).

2.1. Eesmärk

Eesmärk on koguda, töödelda ja analüüsida eesti-inglise koodivahetusega lauseid Redditist. Koodivahetuse puhul mõistetakse käesolevas töös nähtust, kus ühe lause piires esineb kaks erineva keele keeleüksust – antud juhul eesti ja inglise keel. Kogumisel, töötlemisel ja analüüsimisel on kolm omavahel seotud eesmärki:

- luua piisavalt mahukas, kuid samuti mitmekesine korpus analüüsiks;
- valmistada kogutud materjal ette käsitsi märgenduseks;
- võrrelda olemasolevaid, laialdaselt kasutusel olevaid ning projekti jaoks sobivaid keeletuvastajaid ning hinnata nende sobivust eesti-inglise koodivahetuse tuvastamiseks.

Andmekogumise lähtekoht on kasutaja loodud tekst avalikust veebikeskkonnast, millele rakendatakse automaatne filtreerimine ning koodivahetuse esmane tuvastus. Kogu protsess on valminud Pythoni skriptidena, mille erinevad osad on kirjeldatud allpool. Kood ning avatud andmed, välja arvatud tšätikorpus, on saadaval Githubis[20].

2.2. Andmekogumise arhitektuur

Andmete allikaks valiti Redditi alafoorum r/Eesti, mis on suurim eestikeelne alafoorum Redditis. Kuna korpuse jaoks oli vaja tekste eri valdkondadest, siis Reddit sobis selle eesmärgi jaoks hästi. Redditi kasutajad kirjutavad peamiselt vabas vormis, mis võib suurendada tõenäosust, et ingliskeelseid väljendeid või repliike esineb eestikeelse teksti sees selle loomuliku osana. Andmete kogumiseks kasutati Redditi avalikke JSON-i lõpp-punkte (ingl *endpoint*), kasutamata Redditi rakendusliidest (ingl *API*), kuna autoril ei õnnestunud saada sellele ligipääsu. See tähendab, et ligipääs toimub üldkasutatava veebiliidesega. Kogumine toimib kategooriapõhiselt:

[20] https://github.com/MarkoSasi/keeletuvastajate-analyys

kordamööda otsitakse postitusi kategooriatest *hot, new, top* ja *rising*. Kategooria *top* puhul rakendatakse ajalisi filtreid (*day, week, month, year, all*), et kaasata nii värsket kui ka populaarset sisu. Kui kategooriapõhiselt kogutud materjal osutub ebapiisavaks, alustatakse märksõnapõhist kogumist. Märksõnad on jaotatud teemadena, kus võib esineda koodivahetust nagu näiteks tehnoloogia, haridus, tervis, igapäevaelu. Märksõnade täielik loetelu on esitatud lisas 1.
Iga postituse puhul töödeldakse kolme tekstiliiki: postituse pealkiri, postituse sisu ning kommentaarid. Kommentaare eraldatakse rekursiivselt kogu lõimest, mille tulemusena saadakse nii vastused, kui ka alavestlused. Kasutajate ning moderaatorite poolt eemaldatud kommentaarid ([removed]) jäetakse kogumisest välja.
Tehniline töökindlus on tagatud mitmekihilise päringuloogikaga. Päringupiirangu (ingl *Rate limit*) vältimiseks on päringute vahel viivitus 2,5 sekundit koos juhusliku ajalise hajutusega (ingl *jitter)*. Kasutajaagendi tunnus (ingl *User-Agent*) vahetub juhuslikult. Kui server tagastab veakoodi 429 (*Too Many Requests)*, järgib süsteem ooteaega (*Retry-After)*. Maksimaalne korduskatsete arv ühe päringu kohta on kolm. Pikaajalise kogumise jätkuvuse tagamiseks toetab programm kontrollpunktipõhist jätkamist.

2.3. Algne töötlus ja kvaliteedikontroll

Enne analüüsi läbib iga kogutud tekst mitmeastmelise puhastuse. Esimeses etapis eemaldatakse kõik mitte-keelelised osad: URL-id, koodiblokid, Redditi spetsiifilised viited kasutajatele (u/) ja alamfoorumitele (r/). Need ei kanna vajalikku infot ning võivad mõjutada tuvastajate tulemusi.
Teises etapis jagatakse puhastatud tekst lausteteks kirjavahemärkide alusel.
Kolmandas etapis rakendatakse minimaalne lausepikkuse filter: laused, milles on vähem kui viis sõna jäetakse välja, kuna sellised laused on analüüsiks ebapiisavad.
Neljas etapp on duplikaatide eemaldamine, mis tagab, et üks ja sama lause ei esine korpuses mitu korda.

2.4. Koodivahetuse esmase tuvastamise kriteeriumid

Pärast tekstide eeltöötlust rakendatakse korpusele automaatne, skriptipõhine filter, mille eesmärgiks on valida välja edasiseks käsitsi märgendamiseks võimalikult palju koodivahetust sisaldavaid lauseid ning vähendada valepositiivsete vastete hulka. Valepositiivseteks loetakse

lauseid, kus ingliskeelne sõna on tegelikult laensõna, brändinimi või juhuslik sõne, mis ei peegelda koodivahetust.

Lauseid loetakse koodivahetust sisaldavaks, kui täidetud on järgmised tingimused:

- lauses esineb nii eesti- kui ka ingliskeelne eristatav tekstiosa;
- ingliskeelseid sõnu on vähemalt kaks;
- eestikeelseid sõnu on vähemalt kaks;
- Lauses esineb tugev eesti keelele omane järelliide (-nud, -tud, -ks) või erimärk (nt ä, ö, ü, õ);
- Lauses esineb vähemalt üks inglise keelne artikkel *the*, *a*, *an* või eessõna *of*, *in*, *on.*

2.5. Töövoo struktuur

Kogu projekt on jagatud iseseisvateks Pythoni skriptideks, mille ülesanded on eristatud. See võimaldab töövoo korduvkasutust ning teistel isikutel katseid korrata. Skriptid on järgmised:

- src/config.py – haldab parameetreid ja märksõnu;
- src/scraper.py – vastutab Redditi andmete kogumise eest;
- src/sentence_processor.py – tegeleb puhastuse ja lauseteks jagamisega;
- src/language_detector.py – rakendab koodivahetuse tuvastuse;
- src/main.py – koordineerib põhitöövoogu.

2.6. Töövoo piirangud

Siiski tuleb arvestada mitmete piirangutega. Esiteks, kuna andmete kogumiseks kasutatakse Redditi avalikke JSON-i lõpp-punkte ilma ametliku rakendusliideseta, on päringute arv ajaliselt piiratud ning serveripoolsed muutused võivad mõjutada kogumise tõhusust. Teiseks ei vasta r/Eesti alafoorumis olevad tekstid sageli kirjakeele normidele ning sageli kasutatakse lühendeid, mis raskendavad automaatset keeletuvastust. Kolmandaks sõltub koodivahetuse tuvastuse täpsus suuresti kriteeriumite parameetritest, mille optimaalsust on keeruline põhjendada.

2.7. Eetilised kaalutlused

Reddit on avalik platvorm, mille kasutustingimused lubavad avalike postituste teaduslikku analüüsi (Reddit Inc. s. a.). Kogutud andmeid ei seostata isikuandmetega ning analüüsitavaks on

lause, mitte kasutaja. Töös ei avalikustata kasutajanimesid ega muid identifitseerivaid andmeid. Korpuses säilitatakse vaid lauseteks segmenteeritud tekst koos keelelise märgendusega.

2.8. Korpused

Keeletuvastajate hindamiseks kasutati kahte korpust: Tšätikorpus ja Redditi r/Eesti alafoorumist kogutud Redditi korpus. Mõlemad korpused sisaldavad eestikeelset teksti, milles esineb ingliskeelseid sõnesid. Järgnevalt kirjeldatakse mõlema korpuse päritolu, ülesehitust ja põhilisi näitajaid.

2.8.1 Tšätikorpus

Tšätikorpus on osa projekti Teismeliste keel Eestis (TEKE)[21] raames kogutud teismeliste keele korpusest, mis sisaldab privaatseid sõnumivestlusi (peamiselt Facebook Messengeri vestlused) noorte eesti keele kõnelejate vahel (Koreinik jt 2022). Korpus koosneb 100 TSV-failist, millest iga fail vastab ühele vestlusele. Failinimed on märgistatud koodidega, milles esiosa tähistab piirkonda (nt Ant = Antsla, Trt = Tartu), sellele järgnev number ja täht näitavad vanuseklassi ja sugu ning lõpuosa on osaleja tunnuskood. Iga fail sisaldab viit veergu: Jutt (kas rida sisaldab kasutaja teksti), Aeg (millal, sõnum saadeti), Osaleja (anonüümne osaleja tunnus), Puhas_tekst (algne tekst) ja Margendatud_tekst (märgendatud tekst). Ridade filter Jutt = “Jah” eraldab kasutaja saadetud reaalseid vestlussõnumeid.

Tšätikorpuses on kasutusel kolme tüüpi märgendus: keele märgendus (ann_lang), lühendi märgendus (ann_abbr) kolme alamtüübiga (CUT – lühendamine, INIT – initsiaalid, PRON – häälduspõhine lühend) ja konfidentsiaalse info märgendus (ann_conf). Ingliskeelseid sõnesid märgendatakse sõnekaupa sildiga <ann_lang='eng'>sõna</>, lühendite puhul, mis on inglise päritoluga, kasutatakse kombineeritud silti, nt <ann_abbr='INIT'_lang='eng'>lol</>.

Käesoleva töö jaoks võeti korpusest vaid read, mille veeru Jutt väärtus on „jah”, st kasutajate endi saadetud sõnumid. Kokku vastab sellele tingimusele 15 828 rida 18 106-st, millest 14 407 real esineb vähemalt üks sõne. Kogu korpus sisaldab 61 333 sõnet, millest 56 882 (92,74%) on märgendatud eestikeelsetena ja 4451 (7,26%) ingliskeelsetena. Ingliskeelseid sõnesid sisaldavaid sõnumeid on 2212. Sõnumite keskmine pikkus on 4,3 sõnet (mediaan 3 sõnet), mis on tüüpiline

[21] https://sisu.ut.ee/teke/

lühikestele sõnumivestlustele, ning pikim sõnum sisaldab 432 sõnet. Tšätikorpuse tekstid on iseloomult mitteametlikud, sisaldavad rohkesti lühendeid, kõnekeelseid vorme ja mittestandardset ortograafiat, mis muudab korpuse keeletuvastajatele keerukaks testmaterjaliks.

2.8.2. Redditi korpus

Redditi korpus sisaldab eesti-inglise koodivahetusega lauseid Redditi alafoorumist r/Eesti. Andmete kogumise ja töötlemise protsess on täpsemalt kirjeldatud andmekogumise arhitektuuri osas. Korpus koosneb 534 lausest, mis on käsitsi läbi vaadatud ja märgendatud sõnetasandil samade märgenduspõhimõtete järgi, mida kasutatakse Tšätikorpuses (vt peatükk „Märgenduspõhimõtted"). Iga lause ingliskeelsed sõned on märgendatud sildiga <ann_lang='eng'>sõna</>. Sõnad märgendatakse ükshaaval, mitte fraasidena. Märgendatud korpus asub failis reddit_margendus.xlsx, mis sisaldab kahte veergu: Puhas_tekst (algne lause) ja Margendatud_tekst (märgendatud lause).

Korpus sisaldab kokku 9 773 sõnet, millest 7 778 (79,59%) on märgendatud eestikeelsetena ja 1 995 (20,41%) ingliskeelsetena. Kõik 534 lauset sisaldavad vähemalt ühte ingliskeelset sõnet, kuna lausete kogumise protsessis rakendati koodivahetuse filtrit, mis nõudis nii eesti- kui ka ingliskeelsete sõnede olemasolu (vt 3.4). Lausete keskmine pikkus on 18,3 sõnet, mis on Tšätikorpuse sõnumitest oluliselt pikem, pikim lause sisaldab 81 sõnet ja lühim 5 sõnet. Redditi korpuse tekstid on pärit kommentaaridest, mis on stiililt poolformaalsed ehk kasutajad kirjutavad vabas vormis, kuid avaliku suhtluse tõttu on tekstid keskmiselt pikemad, lauseliselt ühtlasemad ja ortograafiliselt korrektsemad kui privaatsed sõnumivestlused. Ingliskeelsete sõnede osakaal (20,41%) on Tšätikorpusega (7,26%) võrreldes oluliselt suurem, mis tuleneb kogumismetoodikast – Redditi korpusesse valiti ainult need laused, milles koodivahetus on automaatselt tuvastatud.

2.8.3. Korpuste võrdlus

Kaks korpust täiendavad üksteist ja võimaldavad hinnata keeletuvastajate töökindlust erinevat tüüpi tekstidel. Tabelis 1 on esitatud mõlema korpuse põhilised arvnäitajad.

Tabel 1. Korpuste põhilised arvnäitajad.

Näitaja	Tšätikorpus	Redditi korpus
Allikas	Messengeri vestlused	r/Eesti (Reddit)
Faile / lauseid	97 faili (15 573 sõnumit)	534 lauset
Sõnesid kokku	61 333	9 773
Eestikeelseid sõnesid	56 882 (92,74%)	7 778 (79,59%)
Ingliskeelseid sõnesid	4451 (7,26%)	1 995 (20,41%)
Sõnumeid/lauseid ingliskeelsete sõnedega	2212 (14,2%)	534 (100%)
Keskmine pikkus (sõnedes)	4,3	18,3
Maksimaalne pikkus	432	81
Teksti tüüp	privaatne, mitteformaalne	avalik, poolformaalne

Tabelist nähtub, et korpused erinevad nii mahu, tekstitüübi kui ka ingliskeelsete sõnede osakaalu poolest. Tšätikorpus on märkimisväärselt suurem (61 333 sõnet vs 9 773 sõnet), kuid ingliskeelsete sõnede osakaal on oluliselt väiksem (7,26% vs 20,41%).

Sõnumite pikkuse erinevus on samuti märkimisväärne: Tšätikorpuse sõnumid on keskmiselt 4,3 sõnet pikad, Redditi laused aga 18,3 sõnet. Keeletuvastajatele tähendab see, et Tšätikorpuse sõnumid pakuvad iga otsuse kohta vähem konteksti, mis muudab ingliskeelsete sõnede tuvastamise raskemaks. (Mitte kõik tuvastajad ei kasuta lausekonteksti.)

Koodivahetus on samuti erinev kahe korpuse lõikes. Esialgse mulje põhjal tundub, et Tšätikorpuses esineb koodivahetus sageli üksikute sõnadena (nt *sorry, lol, ok*), mis on tihtipeale juba eesti keeles juurdunud laenud ning mille keeleline kuuluvus on hägune (Vihman jt 2023). Redditi korpuses avaldub koodivahetus tüüpilisemalt pikemate fraasidena (nt *this is really cool*).

Mõlema korpuse märgendusskeem on ühildatav, mis võimaldab keeletuvastajaid hinnata samadel põhimõtetel. Märgenduse ühilduvuse ja kohandamise üksikasjad on esitatud peatükis „Märgenduspõhimõtted".

3. Märgenduspõhimõtted

Redditi korpuse käsitsi märgendamisel võeti aluseks Tšätikorpuse märgendusjuhend. Tšätikorpuses kasutatakse kolme märgendustüüpi: konfidentsiaalse info märgendus (*ann_conf*), keele märgendus (*ann_lang*) ning lühendi märgendus (*ann_abbr*) kolme alamtüübiga: lühendamine (CUT), initsiaalid (INIT) ja häälduspõhine lühend (PRON). Ingliskeelsed sõnad märgendatakse sõnahaaval, st iga sõna saab eraldi sildi <*ann_lang='eng'>sõna*</>, mitte terve fraas korraga. Kuna Redditi korpus erineb Tšätikorpusest sisu poolest (tegemist on avalike foorumipostitustega, mitte privaatsete sõnumivestlustega) oli vaja mõningaid märgenduspõhimõtteid kohandada.

3.1. Nimed ja ettevõtete nimed

Tšätikorpuse märgendusjuhendi kohaselt nimesid keele järgi ei märgendata (*„Names were not coded for language")* (Vihman jt 2023). Redditi korpuses järgiti sama põhimõtet: ettevõtte- ja tootenimed nagu *Google*, *TikTok*, *Reddit*, *Spotify*, *Caddy*, *Corolla* jt ei saa keele märgendit, isegi kui need on ingliskeelse päritoluga. Samuti ei kasutata konfidentsiaalsuse märgendit (*ann_conf*), kuna Redditi postitused on avalikud.

Ettevõtte- ja üldnime vahel tehakse siiski vahet. Ettevõttenimi ise jääb märgendamata, kuid sellest tuletatud üldnimi on ingliskeelne sõna ja märgendatakse vastavalt. Näiteks *YouTube* on ettevõttenimi ja jääb märgendamata, kuid *youtuber* on tuletatud üldnimi, mida kasutatakse kui ingliskeelset üldnime, mistõttu saab see märgendi. Sama loogika kehtib näiteks sõnavormidele *echochamber* ja *downvotesi*, mis on tuletatud platvormispetsiifilisest sõnavarast, kuid toimivad lauses üldnimedena.

3.2. Liitsõnad ja segavormid

Redditi tekstides esineb sõnu, kus ingliskeelsele tüvele on liidetud eesti keele muutelõpp, sageli sidekriipsuga (nt *data-transferit*, *example-it*, *race-i*). Need ei ole liitsõnad, vaid ingliskeelsed sõnad eesti keele käände- või pöördelõpuga. Tšätikorpuses sellised vormid praktiliselt ei esine, mistõttu ei saa selle märgendust eeskujuna kasutada. Käesoleva töö jaoks otsustati märgendada terve sõna ingliskeelsena, st <*ann_lang='eng'>kohtucase-i*</>. Sõna sisaldab ingliskeelset tüve,

mis on koodivahetuse olemuslik osa, ning sõna tükeldamine osadeks muudaks märgendussüsteemi ebaühtlaseks ja raskesti töödeldavaks.

3.3. Lühendid

Ingliskeelsed lühendid ja akronüümid märgendatakse kombineeritud märgendiga *<ann_abbr='INIT'_lang='eng'>*, mis näitab nii lühendi tüüpi kui ka keelt. See vastab Tšätikorpuse näidetele, kus *lol*, *omg*, *wtf* ja *btw* on märgendatud samal viisil. Redditi korpuses laiendati seda põhimõtet ka tehnilistele lühenditele: *API*, *GPS*, *USB*, *AI*, *IT*, *CV*, *LLM* jt said samuti INIT-sildi koos keele märgendusega. Kuigi mõned neist (nt *GPS*, *IT*) on eesti keeles väga levinud, otsustati järjepidevuse huvides märgendada kõik ingliskeelse päritoluga akronüümid ühtemoodi. Alternatiivina oleks saanud rahvusvahelised lühendid jätta märgendamata, kuid see oleks tekitanud küsimuse: millisest hetkest on lühend piisavalt "eestikeelestunud", et seda enam mitte märgendada.

3.4. Morfoloogiliselt kohandatud sõnad ja laensõnade piir

Redditi tekstides esineb rohkesti ingliskeelseid sõnu, millele on lisatud eesti käände- ja pöördelõppe, nt *followisin*, *retweetisid*, *downvotesi*, *supportib*. Tšätikorpuses märgendatakse sellised sõnad ingliskeelsetena (nt *leftisin*, *editisid*, *accounti*) ning Redditi korpuses järgiti sama põhimõtet.

Eraldi kaalumist vajavad sõnad, mis on küll inglise päritoluga, kuid juba eesti keeles juurdunud. Mitmed sellised sõnad esinevad ka Sõnaveebis[22], näiteks *skoorima*, *feilima*, *bugi* ja *mega*. Nende puhul tekib küsimus, kas tegemist on koodivahetusega või juba laensõnaga.

Käesolevas töös otsustati sellised sõnad siiski märgendada ingliskeelsetena. Kuigi need esinevad Sõnaveebis, on nende ingliskeelne päritolu kõnelejale tõenäoliselt äratuntav ning need erinevad tavapärastest eestikeelsetest sõnadest. Ka Tšätikorpuses on analoogsed sõnad nagu *sorry* (mis on samuti Sõnaveebis olemas) järjepidevalt märgendatud ingliskeelsetena, samas sõna *mega* on Tšätikorpuses jäetud märgendamata. Tuleb tunnistada, et piir laensõna ja koodivahetuse vahel on paratamatult hägune ning teistsugune otsus oleks samuti põhjendatav.

[22] https://sonaveeb.ee/

3.5. Kokkuvõte märgendusreeglitest

Kokkuvõtlikult järgib Redditi korpuse märgendamine Tšätikorpuse süsteemi järgmiste täpsustustega. Nii Tšätikorpuses kui ka Redditi korpuses jäävad nimed (sealhulgas ettevõtte- ja tootenimed) keele järgi märgendamata. Redditi korpuses täiendab seda põhimõtet üksnes selgesõnaline eristus nime ja sellest tuletatud üldnimeks kasutatava sõnavormi (nt *Youtube* ja *youtuber*) vahel, kus viimane saab ingliskeelse märgendi. Ingliskeelsed lühendid saavad kombineeritud sildi *ann_abbr='INIT'_lang='eng'*; segaliitsõnad märgendatakse tervikuna ingliskeelsetena; morfoloogiliselt kohandatud ingliskeelsed sõnad märgendatakse ingliskeelsetena ka juhul, kui sõnatüvi on Sõnaveebis registreeritud. Konfidentsiaalsuse märgendit ei kasutata, kuna Redditi postitused on avalikud.

4. Eksperimendid

4.1. Töövoog

Töövoog on jaotatud kaheks eraldiseisvaks osaks. Esimene osa tegeleb Redditi andmete kogumise, esmase filtreerimise ning märgendamiseks ettevalmistamisega.

Teine osa viib läbi keeletuvastajate hindamise juba käsitsi märgendatud Tšätikorpusel ja Redditi korpusel. Hindamiskood paikneb kaustas *evaluation/* ning peamiseks hindamisskriptiks on *full_identifier_report.py*, mille väljundit on kasutatud tulemustabelite koostamiseks. Selline jaotus tagab, et andmete kogumine ning tulemuste hindamine ei satu üksteisega vastuollu ning kumbki tööosa on iseseisvalt rakendatav.

Töövoo skriptide kirjutamiseks kasutati vähesel määral tehisaru rakendust Claude Code (Opus 4.7) (Anthropic 2026), eelkõige üksikute Pythoni funktsioonide ja kommentaaride sõnastamiseks. Kogu kood on autori poolt üle vaadatud, testitud ja vajadusel kohandatud. Tehisaru rakendust ei kasutatud käesoleva töö teksti või analüüsi koostamiseks.

Sisendina kasutatakse Tšätikorpuse puhul ainult sõnumeid, kus veeru *Jutt* väärtus on *jah* (st kasutaja enda saadetud sõnumid) ja milles esineb vähemalt üks ingliskeelne sõne. Redditi korpuse puhul kasutatakse kogu märgendatud andmestikku failist *reddit_margendus.xlsx*.

4.2. Sisend

Kasutatud keeletuvastajad erinevad üksteisest selle poolest, milline on nende sisend (üksiksõna, lause või terviklik tekst) ja kas nad tagastavad ühe keelemärgendi terve sisendi kohta või eristavad lauses mitut keelt. Selleks, et neid kõiki saaks võrrelda, jaotati nad kahte rühma.

Esimene ja suurem rühm on sõnetasandi tuvastajad, kuhu kuuluvad heliport, glotlid, nb-nordic-lid, fastText, fast-langdetect, pycld2, langid, lingua ja masklid. Need tuvastajad määravad kogu sisendile ühe keele. Käesolevas töös rakendatakse neid sõne kaupa, st iga sõne on eraldi sisend. Tasub silmas pidada, et need tuvastajad ei kasuta sellise rakenduse korral lausekonteksti. Naabersõnad ei jõua mudelini ning see vähendab täpsust.

Teine rühm on lausetasandi tuvastajad, kuhu kuuluvad AnE-LID, lingua-mixed ja masklid-cs. Need mudelid toetavad mitme keele tuvastamist ja neile antakse sisendiks terve lause korraga. Mudel leiab lause seest ise ingliskeelsed fraasid.

4.3. Hindamine

Hindamise arvutamine (täpsus, saagis, õigsus ja F1-skoor) on leitav failis *evaluation/full_identifier_report.py* ning see tugineb Pythoni teegile *scikit-learn (*Pedregosa jt 2011). Kasutusel on järgnevad funktsioonid: *accuracy_score, classification_report, confusion_matrix, f1_score* ja *precision_recall_fscore_support*. Tulemustabelis esitatud tulemused tuletatakse otseselt nende funktsioonide väljunditest. Koondskoor on nende keskmine tulemus.

Iga lause korral võrreldakse iga tuvastatud sõne puhul tuvastaja märgendit sõne kuldmärgendiga. Sõned, mida kuldstandardi järgi ingliskeelseks ei märgendata, loetakse hindamises eestikeelseteks.

4.4. Väljund

Hindamise põhiskript *full_identifier_report.py* salvestab tulemused XLSX-vormingus – Tšätikorpuse kohta faili *chat_tsv_full_report.xlsx* ja Redditi korpuse kohta faili *reddit_full_report.xlsx*. Mõlemad failid sisaldavad kahte töölehte: *SentenceView* ning *Summary.* Töölehes *SentenceView* on esitatud tulemused lause kaupa. Üks rida vastab ühele lausele või sõnumile. Veerg *id* on järjekorranumber, *file* viitab lähtefailile ja *sentence* sisaldab lause teksti. Veerg *gold_cs_parts* näitab kuldstandardi järgi ingliskeelseks märgendatud sõnesid. Iga keeletuvastaja leitud ingliskeelsed sõned on tabelis märgitud veeru*l (tuvastaja)_cs_parts*. Lõigud on eraldatud püstkriipsuga (|) ning mõttekriips (-) tähistab, et tuvastaja ei leidnud lauses ühtegi ingliskeelset sõne.

Tööleht *Summary* koondab iga tuvastaja tulemused kokku. Veerud on *identifier* (tuvastaja nimi), *token_accuracy_pct* (sõne õigsus), *main_language_est_accuracy_pct* (eestikeelse sõne täpsus), *eng_token_precision_pct*, *eng_token_recall_pct* ja *eng_token_f1_pct* (ingliskeelsete sõnede täpsus, saagis ja F1-skoor) ning *composite_score_pct* (koondskoor). Read on järjestatud koondskoori järgi kahanevalt. Nende mõõdikute põhjal on koostatud peatükis 5. esitatud tulemustabelid.

Töölehe *SentenceView* ülesehitust illustreerib järgmine näide Redditi korpusest. Kuldmärgenduse järgi on ingliskeelseks määratud fraas *it’s a numbers game* ning iga tuvastaja real on näha, millise lõigu vastav tuvastaja samast lausest leidis.

Lause: Ole valmis neid kirju palju saatma it's a numbers game

Kuldstandard: it's a numbers game

ane-lid: it's a numbers game

lingua-mixed: numbers game

heliport: it's | numbers game

langid: Ole | neid | palju | it's | numbers game

Tuvastajad käitusid selle lausega erinevalt. Tuvastaja AnE-LID leidis täpselt sama lõigu, mille märkis kuldstandard. Tuvastaja lingua-mixed leidis koodivahetuse üles, kuid jättis lõigu algusest sõnad it's a vahele. Tuvastaja heliport jagas ingliskeelse osa kaheks eraldi lõiguks. Tuvastaja langid märgendas ingliskeelseks ka eestikeelsed sõnad Ole, neid ja palju; need on valepositiivsed vasted. Sellised erinevused tuvastajate vahel koondab tööleht *Summary* arvulisteks mõõdikuteks.

5. Tulemused

5.1. Hindamiskriteeriumid

Keeletuvastajate hindamiseks kasutati viit parameetrit, mis katavad erinevaid koodivahetuse tuvastuse nähtusi. Sõne õigsus (ingl *token accuracy*) näitab, kui suure osa kõikidest sõnedest tuvastaja õigesti märgendas, olenemata sellest, kas need olid eesti- või ingliskeelsed. Eestikeelse sõne täpsus mõõdab, kui täpselt tuvastati eestikeelsed sõned. Ingliskeelsete sõnede hindamisel kasutati kolme parameetrit: täpsust (*precision*), mis näitab, kui suur osa ingliskeelseks märgendatutest olid tegelikult ingliskeelsed; saagist (*recall*), mis näitab, kui suure osa tegelikult ingliskeelsetest sõnedest tuvastaja leidis ning F1-skoori, mis ühendab need kaks koondskooriks. Koondskoor on kõigi skooride keskmine, mis annab ühtse ülevaate tuvastaja üldisest tulemusest.

5.2. Tšätikorpuse tulemused

Tšätikorpuse tulemused (vt tabel 2) toovad selgelt esile erinevused koodivahetusele spetsialiseeritud ning üldiste keeletuvastajate vahel. Parima koondskoori saavutas heliport (74,32%), edestades teisele kohale tulnud AnE-LIDi (71,49%) 2,83 protsendipunktiga. Kolmandale kohale tuli glotlid (68,9%). Lingua mudelid lingua-mixed (64,53%) ja lingua (64,27%) jäid neljandaks ja viiendaks.

Konservatiivsete tuvastajate rühma kuuluvad heliport (74,32%), glotlid (68,9%) ja nb-nordic-lid (60,54%). Kuigi heliport on koondskoori järgi Tšätikorpuse parim, on tema käitumismuster (kõrge täpsus, madal saagis) ühtne ülejäänud konservatiivsete tuvastajatega. Need saavutasid kõrge ingliskeelse täpsuse, kuid jäid saagise poolest selgelt alla. Heliporti ingliskeelse sõne määramise täpsus oli 87,08%, glotlidil 54,28%, kuid mõlema saagis jäi alla 36% (heliport 35,12%, glotlid 30,08%). Vastupidist mustrit demonstreerisid masklid-cs, masklid ja langid, mille ingliskeelse sõne määramise saagis ulatus 85%-st 91%-ni (masklid 84,97%, masklid-cs 89,82%, langid 90,27%), kuid täpsus jäi alla 15% (vastavalt 13,08%, 13,13% ja 8,62%). Langidi koondskoor (24,57%) oli madalaim ka Redditi korpusel (42,69%).

Tabel 2. Keeletuvastajate tulemused Tšätikorpusel (kasutatud ainult eesti-inglise koodivahetust sisaldavaid sõnumeid).

Tuvastaja	Sõne õigsus (%)	Eesti sõne täpsus (%)	Ingl. täpsus (%)	Ingl. saagis (%)	Ingl. F1 (%)	Koond (%)
heliport	94,99	99,6	87,08	35,12	50,05	74,32
ane-lid	90,62	91,04	42,28	85,15	56,5	71,49
glotlid	93,19	98,05	54,28	30,08	38,71	68,9
lingua-mixed	86,87	87,36	32,9	80,45	46,7	64,53
lingua	87,43	88,62	32,76	72,01	45,03	64,27
nb-nordic-lid	89,33	93,87	27,62	30,4	28,94	60,54
fast-langdetect	81,38	82,76	22,09	63,45	32,77	55,67
fasttext	80,98	82,44	21,4	62,08	31,82	55,35
pycld2	76,89	77,34	19,46	71,11	30,56	52,59
masklid	58,53	56,5	13,08	84,97	22,67	40,3
masklid-cs	56,76	54,21	13,13	89,82	22,9	39,38
langid	30,89	26,31	8,62	90,27	15,74	24,57

5.3. Redditi korpuse tulemused

Redditi korpust testiti samade tuvastajate peal, mis Tšätikorpuse puhul (vt tabel 3). AnE-LID oli taas parim tuvastaja koondskoori poolest (93,49%), olles märkimisväärselt parem ka Tšätikorpuse tulemusest. Ingliskeelse sõna F1-skoor ulatus 91,67%-ni, mis näitab, et tuvastaja töötab Redditist kogutud lausetel hästi.

Lingua-mixed tuli teisele kohale (87,63%). Esile tuleks tuua ka heliporti tulemus – eesti sõne täpsus oli 99,68%, mis on kõrgeim kõigi tuvastajate seas, ning ingliskeelse sõne määramise täpsus 97,43%. Samas jäi heliporti ingliskeelse sõne määramise saagis väga madalaks (47,47%), mis tähendab, et mudel jätab ligi poole ingliskeelsetest sõnedest tuvastamata. Sama muster esineb ka glotlidi (saagis 37,49%) ja nb-nordic-lidi (saagis 46,67%) puhul. Need tuvastajad on konservatiivsed ehk välditakse valepositiivseid ja märgendatakse kahtlased sõned pigem eestikeelseteks.

Langid osutus mõlemal korpusel halvimaks tuvastajaks – Redditi korpusel jäi selle ingliskeelse sõne määramise täpsus vaid 27,01% ning eesti sõne täpsus 34,79%, mis tähendab, et tuvastaja ei suuda eesti- ja ingliskeelseid sõnu üksteisest piisavalt hästi eristada.

Tabel 3. Keeletuvastajate tulemused Redditi korpusel.

Tuvastaja	Sõne õigsus (%)	Eesti sõne täpsus (%)	Ingl. täpsus (%)	Ingl. saagis (%)	Ingl. F1 (%)	Koond (%)
ane-lid	96,4	96,22	86,82	97,09	91,67	93,49
lingua-mixed	92,37	94,66	80,04	83,41	81,69	87,63
heliport	89,02	99,68	97,43	47,47	63,84	84,92
glotlid	86,42	98,97	90,34	37,49	52,99	82,26
lingua	90,38	92,02	72,96	84,01	78,1	80,85
nb-nordic-lid	85,27	95,17	71,23	46,67	56,39	76,55
pycld2	84,97	84,93	59,16	85,11	69,8	71,82
fast-langdetect	83,68	85,78	57,66	75,49	65,38	71,03
fasttext	83,63	85,7	57,54	75,54	65,32	70,89
masklid-cs	72,69	68,08	42,15	90,68	57,55	59,31
masklid	72,62	68,09	42,05	90,28	57,37	59,24
langid	46,89	34,79	27,01	94,09	41,97	42,69

5.4. Tuvastajate võrdlev analüüs

Korpuste tulemuste võrdlemine toob esile seaduspärasusi. Esiteks oli kõikide kaheteistkümne tuvastaja koondskoor Redditi korpusel kõrgem kui Tšätikorpusel: AnE-LIDi puhul 22,00 protsendipunkti võrra (71,49%'st 93,49%'ni), heliporti puhul 10,60 protsendipunkti võrra (74,32%'st 84,92%'ni) ja glotlidi puhul 13,36 protsendipunkti võrra (68,9%'st 82,26%'ni). Erinevus tuleneb suure tõenäosusega kahe korpuse iseloomulikest erinevustest: Tšätikorpus sisaldab privaatseid sõnumivestlusi, milles esineb rohkem lühendeid, kõnekeelseid vorme ja lühikesi väljendeid, samal ajal kui Redditi avalikud postitused on keskmiselt pikemad ja õigekirjaliselt ühtlasemad.

Teine järjepidev muster on tuvastajate jagunemine kolmeks rühmaks täpsuse ja saagise järgi. Esimese rühma moodustavad koodivahetust hästi tuvastavad tuvastajad AnE-LID, lingua-mixed ja lingua, mille puhul ingliskeelse täpsuse ja saagise vahe on suhteliselt väike. Teise rühma kuuluvad konservatiivsed tuvastajad heliport, glotlid ja nb-nordic-lid, mis saavutavad väga kõrge täpsuse, kuid jätavad olulise osa ingliskeelsetest sõnedest tuvastamata. Kolmanda rühma moodustavad agressiivsed tuvastajad masklid, masklid-cs ja langid, mille saagis on väga kõrge, kuid täpsus madal. Need kalduvad eestikeelseid sõnu liiga tihti ingliskeelseteks märgendama. Järgnevalt tuuakse iga rühma tuvastajate tugevused ja miinused detailsemalt välja.

5.4.1. Koodivahetusele spetsialiseeritud tuvastajad

AnE-LID saavutas Redditi korpusel parima koondtulemuse (93,49%) ja Tšätikorpusel teise koha (71,49%, vahetult heliporti järel) ning näitab kõige tasakaalustatumat täpsuse ja saagise suhet. Tuvastaja tugevus seisneb selles, et XLM-RoBERTa-large keelemudeli peenhäälestus sõnatasandi koodivahetuse ülesandele võimaldab arvestada nii morfoloogiliselt kohandatud sõnavormidega (nt followisin) kui ka segaliitsõnadega. Redditi korpusel saavutas AnE-LID ingliskeelsete sõnade tuvastamisel F1-skoori 91,67%, mis tähendab, et peaaegu kõik koodivahetuslaused tuvastatakse koodivahetust sisaldavateks. Oluline oleks välja tuua asjaolu, et XLM-RoBERTa-large on muudest mudelitest oluliselt aeglasem.

Masklid ja masklid-cs tulemused on omavahel peaaegu identsed (Tšätikorpusel 40,3% ja 39,38%, Redditi korpusel 59,24% ja 59,31%), kusjuures koodivahetuse režiim predict_codeswitch ei anna olulist eelist alusmeetodi (ingl default method) ees. Mõlema mudeli ingliskeelsete sõnade tuvastamise saagis on küll kõrge (üle 84%), kuid täpsus jäi Tšätikorpusel u 13% piiresse (13,08%

ja 13,13%) ning Redditi korpusel u 42% piiresse (42,05% ja 42,15%). Kuigi MaskLID on välja töötatud koodivahetuse tuvastamiseks, jäi selle tulemus eesti-inglise andmestiku puhul märkimisväärselt madalamaks, kui Kargaran jt (2024) on välja toonud teiste keelepaaride (türgi-inglise, hindi-inglise, nepali-inglise, baski-hispaania) puhul.

5.4.2. Lingua

Lingua ja lingua-mixed pakuvad mõlemad tasakaalustatud tulemust ilma närvivõrkudele tuginemata. Lingua-mixed osutus selgelt tugevamaks variandiks (64,53% Tšätikorpusel ja 87,63% Redditi korpusel), saavutades Redditi korpusel teise koha. Selle peamiseks tugevuseks on koodivahetuse režiim, mis on välja töötatud spetsiaalselt mitmekeelse teksti segmenteerimiseks. Ingliskeelsete sõnede tuvastamise F1-skoor ulatus Redditi korpusel 81,69%-ni, mis on lähedal AnE-LIDi tasemele. Tavaline lingua režiim jäi koondskoori poolest pisut madalamaks (64,27% ja 80,85%), kuid edestas siiski enamikku teisi keeletuvastajaid. Lingua oluline eelis on kerge rakendatavus.

5.4.3. Konservatiivsed tuvastajad

Heliport, glotlid ja nb-nordic-lid moodustavad selgelt eristuva rühma, mille käitumismuster on üksteisele väga sarnane - kõrge ingliskeelsete sõnede tuvastamise täpsus ja madal saagis. Heliport saavutas mõlemal korpusel kõrgeima ingliskeelse täpsuse (87,08% ja 97,43%), kuid saagis oli (35,12% ja 47,47%). Vaatamata konservatiivsele käitumisele saavutas heliport Tšätikorpuse parima koondskoori, sest selle väga kõrge eesti sõne täpsus tasakaalustab madalat ingliskeelset saagist. Glotlid näitas sarnast mustrit (saagis vastavalt 30,08% ja 37,49%), samuti nb-nordic-lid (30,4% ja 46,67%). Need tuvastajad jätavad seega märkimisväärse osa ingliskeelsetest sõnedest tuvastamata, määrates need pigem eestikeelseteks. Põhjus on selles, et need mudelid on välja töötatud terviku- või lausetasandi keele tuvastamiseks ning eelistavad määrata kogu sisendile selle domineeriva keele märgendi. Lühikeste ingliskeelsete üksiksõnade tuvastamine eestikeelses kontekstis pole nende mudelite peamine ülesanne.

Heliporti tugevuseks on kõrge eestikeelse sõne täpsus (99,6% ja 99,68%) ning kiirus, kuna Rust programmeerimiskeelne teostus on Java versioonist umbes 25 korda kiirem (Zaragoza-Bernabeu 2024). Glotlid katab 1832 keelt ja sobib eelkõige vähese ressurssidega keelte tuvastamiseks (Kargaran jt 2023), kuid eesti-inglise koodivahetuse jaoks pole see optimaalne. Nb-nordic-lid on

välja töötatud Põhja-Euroopa keelte tarbeks, ent piiratud avalik dokumentatsioon ja keskmine koondskoor ei tee sellest käesoleva ülesande puhul tugevat kandidaati.

5.4.4. fastText-il põhinevad tuvastajad

Fasttext ja fast-langdetect tagastavad sisuliselt identseid tulemusi (Tšätikorpusel 55,35% ja 55,67%, Redditi korpusel 70,89% ja 71,03%), mis on ootuspärane: fast-langdetect on optimeeritud Pythoni liides täpselt sama *lid.176* mudeli ümber. Mõlema tuvastaja tulemused jäävad keskmistele kohtadele. Nende tugevuseks on kõrge kiirus, kuid need pole välja töötatud koodivahetuse tuvastamiseks. F1-skoor jäi alla 70% mõlemal korpusel ning ingliskeelse sõne määramise täpsus alla 60%. Fast-langdetecti dokumentatsioonis viidatud kuni 80. kordne kiirendus võib olla teatud juhtudel piisav põhjus eelistamaks seda fastTexti üle.

5.4.5. pycld2

Pycld2 tulemused olid samuti keskmised, kuid Redditi korpusel märksa paremad kui Tšätikorpusel (52,59%'st 71,82%'ni). Tuvastaja eelis on võime tagastada ühe sisendi kohta kuni kolm keelt koos osakaaludega, mis võib teatud kasutusjuhtudel olla väärtuslik. Samas jäi ingliskeelse sõne määramise täpsus mõlemal korpusel madalaks (19,46% ja 59,16%), mis viitab sellele, et CLD2 Naive Bayes klassifikaator ei suuda eesti-inglise koodivahetuses inglise keelt piisava usaldusväärsusega eristada. Saagis on seevastu suhteliselt kõrge (71,11% ja 85,11%), mis paigutab pycld2 lähemale agressiivsete tuvastajate rühmale.

5.4.6. langid

Langid jäi mõlemal korpusel selgelt halvimaks tuvastajaks (24,57% ja 42,69%). Sõne õigsus oli Tšätikorpusel vaid 30,89% ja eesti sõne täpsus 26,31%. Ingliskeelsete sõnede tuvastamise saagis oli Tšätikorpusel kõrgeim kõigist mudelitest (90,27%) ja Redditi korpusel teiseks kõrgeim (94,09%, järgnedes AnE-LIDile), kuid see tuli täpsuse arvelt, mis jäi alla 30% mõlemal korpusel. Tegemist on 2012. aastal välja töötatud mudeliga, mis on välja õpetatud terviktekstide klassifitseerimiseks ja millel pole eraldi mehhanismi koodivahetuse käsitlemiseks. Kuigi langid.py on ajalooliselt olnud üks laialdasemalt kasutatud keeletuvastajaid (Lui & Baldwin 2012), ei ole see eesti-inglise koodivahetuse tuvastamise ülesande jaoks sobiv.

5.5. Soovitused

Läbiviidud katsete tulemustena saab soovitada eesti-inglise koodivahetuse tuvastamiseks keeletuvastajaid erinevateks kasutuskontekstideks. Soovitused lähtuvad sellest, et lõppkasutaja vajadus ning huvi on tuvastada ingliskeelseid sõnesid eestikeelses tekstis võimalikult täpselt ja võimalikult suures osas.

Kui prioriteediks on tuvastuse üldine kvaliteet, on parim valik AnE-LID. Tuvastaja saavutab Redditi korpusel parima ja Tšätikorpusel teise koondtulemuse (Tšätikorpusel edestab teda heliport, kelle tugevus on aga eelkõige väga kõrge eesti sõne täpsus, mitte tasakaalustatud koodivahetuse tuvastus). AnE-LID näitab tasakaalustatud käitumist nii täpsuse kui ka saagise osas. Eriti hästi sobib ta internetitekstile, kus koodivahetus avaldub tihti morfoloogiliselt kohandatud sõnavormide või segaliitsõnadena.

Tugevaim alternatiiv on ka lingua-mixed, mis saavutab teise koha mõlemal korpusel. Lingua-mixed sobib eriti hästi juhul, kui prioriteediks on koodivahetuse saagis, st soov leida võimalikult palju koodivahetust sisaldavaid lauseid.

Kui ülesanne nõuab võimalikult kõrget täpsust, näiteks koodivahetuse märgendamiseks korpuses ilma inimesepoolse järelkontrollita, on otstarbekas valida heliport. See annab erakordselt vähe valepositiivseid tulemusi (ingliskeelsete sõnede tuvastamise täpsus Tšätikorpusel 87,08% ja Redditi korpusel 97,43%), kuid tuleb arvestada, et tuvastatuks osutub vaid osa kogu tegelikust koodivahetusest.

Üldisemate kasutusjuhtude jaoks (nt kiired eelanalüüsid suurtes andmekogudes, kus saagise ja täpsuse vahekord pole esmatähtis) võivad sobida fasttext või fast-langdetect, mis pakuvad rahuldavaid tulemusi. Samas tuleb arvestada, et need pole välja töötatud koodivahetuse tuvastamiseks ja sõnatasandi täpsus jääb spetsialiseeritud lahendustest madalamaks.

Eesti-inglise koodivahetuse tuvastamiseks ei sobi käesolevate tulemuste põhjal langid, masklid ega masklid-cs, mille koondskoor jäi mõlemal korpusel alla 65%. Need tuvastajad kalduvad eestikeelseid sõnu liigselt ingliskeelseteks märgendama, mis vähendab oluliselt nende kasutusväärtust. Kuigi MaskLID on välja töötatud spetsiaalselt koodivahetuse tarbeks, jäid selle tulemused eesti-inglise andmestikul selgelt alla teistele mudelitele.

Kokkuvõtlikult näitavad tulemused, et koodivahetuse sõnatasandi tuvastamine eesti-inglise tekstis on jõukohane ülesanne juhul, kui valitakse selleks spetsiaalselt välja töötatud või vähemalt

mitmekeelse teksti jaoks kohandatud keeletuvastaja. Üldised lausetasandi tuvastajad, isegi tehnoloogiliselt arenenumad, ei suuda asendada koodivahetuse jaoks loodud lahendusi.

Kokkuvõte

Käesoleva töö eesmärk oli võrrelda kahteteist olemasolevat ja laialdaselt kasutatavat keeletuvastajat ning hinnata nende sobivust sõnatasandil eesti-inglise koodivahetuse tuvastamiseks. Eesmärgi saavutamiseks koguti, töödeldi ning märgendati eesti-inglise koodivahetusega lausete korpus Redditi alafoorumist r/Eesti, mida kasutati hindamisel kõrvuti olemasoleva Tšätikorpusega. Töö hõlmas omavahel seotud kolme eesmärki: piisavalt mahuka ja mitmekesise korpuse loomine analüüsiks, kogutud materjali ettevalmistamine käsitsi märgendamiseks ning valitud keeletuvastajate võrdlemine kahel iseloomult erinevalt korpusel koos soovitustega erinevateks kasutuskontekstideks.

Tuvastajaid hinnati kahel korpusel – Tšätikorpus, mis sisaldab 61 333 sõnet privaatsetest Facebook Messengeri sõnumivestlustest ning milles 7,26% sõnedest on märgendatud ingliskeelseks, ning Redditi alafoorumist r/Eesti kogutud ja märgendatud korpus, mis sisaldab 534 lauset ja kokku 9 773 sõnet, millest 20,41% on ingliskeelsed. Mõlema korpuse märgendamisel järgiti Tšätikorpuse märgendusskeemi, mida Redditi eripärasid arvestades vähesel määral kohandati. Tuvastajaid hinnati nelja mõõdiku põhjal – sõne õigsus, täpsus, saagis ja F1-skoor. Hindamiseks kasutati Pythoni teeki scikit-learn.

Parima koondtulemuse saavutas Redditi korpusel AnE-LID (93,49%), edestades selgelt teisi tuvastajaid; Tšätikorpusel oli parim heliport (74,32%), AnE-LID jäi teiseks (71,49%). Teise koha saavutas Redditi korpusel lingua-mixed (87,63%); Tšätikorpusel jäi lingua-mixed neljandaks (64,53%). Tuvastajad jagunesid käitumismustri poolest kolmeks rühmaks. Tasakaalustatud rühma (AnE-LID, lingua-mixed, lingua) kuuluvad mudelid saavutasid ingliskeelsete sõnade tuvastamisel täpsuse ja saagise osas ühtlaselt häid tulemusi. Konservatiivne rühm (heliport, glotlid, nb-nordic-lid) saavutas väga kõrge täpsuse ingliskeelsete sõnade tuvastamisel (heliporti puhul Tšätikorpusel 87,08% ja Redditi korpusel 97,43%) ja ka väga kõrge eestikeelse sõne tuvastamise täpsuse, kuid jäi saagise poolest madalaks ehk oluline osa ingliskeelsetest sõnedest jäid tuvastamata. Agressiivse rühma (masklid, masklid-cs, langid) ingliskeelsete sõnede tuvastamise saagis oli kõrge (langidi puhul ulatus see 90,27%-ni Tšätikorpusel ja 94,09%-ni Redditi korpusel), kuid täpsus jäi mõlemal korpusel alla 30%, mistõttu nende koondtulemus jäi kõige nõrgemaks (langidi puhul 24,57% ja 42,69%).

Kõikide kaheteistkümne tuvastaja koondskoor oli Redditi korpusel kõrgem kui Tšätikorpusel. Erinevus tuleneb suure tõenäosusega kahe korpuse iseloomulikest erinevustest. Tšätikorpuse

privaatsed sõnumivestlused sisaldavad rohkem lühendeid, kõnekeelseid vorme ja lühikesi väljendeid (keskmine sõnumi pikkus 4,3 sõnet), samal ajal kui Redditi avalikud postitused on keskmiselt pikemad (18,3 sõnet lause kohta) ja ortograafiliselt ühtlasemad. See vastab varasemate uuringute tulemustele, mille järgi keeletuvastuse täpsus langeb järsult lühikeste ja mittestandardsete tekstide puhul (Lui & Baldwin 2014; Jauhiainen jt 2019).

Tulemustest lähtuvalt sobib eesti-inglise koodivahetuse sõnatasandi tuvastamiseks kõige paremini üldjuhul AnE-LID, mis on tasakaalustatud nii ingliskeelsete sõnede tuvastamise täpsuse kui ka saagise poolest. Tugevaks alternatiiviks on ka lingua-mixed. Kui ülesanne nõuab võimalikult kõrget täpsust (näiteks koodivahetuse näidete automaatseks lisamiseks korpusesse ilma manuaalse kontrollita), on mõtekas valida heliport, mis annab erakordselt vähe valepositiivseid tulemusi, kuid tuvastab vaid osa kogu tegelikust koodivahetusest. Üldisemate kasutusjuhtude jaoks (näiteks kiired eelanalüüsid suurtes andmekogudes, kus saagise ja täpsuse vahekord pole esmatähtis) sobivad ka fastText ja fast-langdetect. Käesolevate tulemuste põhjal ei sobi eesti-inglise koodivahetuse tuvastamiseks langid, masklid ega masklid-cs, mille koondskoor jäi mõlemal korpusel alla 65% ning mis kalduvad eestikeelseid sõnu liigselt ingliskeelseteks märgendama.

Kokkuvõtlikult on eesti-inglise koodivahetuse sõnatasandi tuvastamine jõukohane ülesanne, kui valitakse selleks spetsiaalselt välja töötatud või vähemalt mitmekeelse teksti jaoks kohandatud keeletuvastaja. Üldised lausetasandi tuvastajad, isegi tehnoloogiliselt arenenumad, ei suuda asendada koodivahetuse jaoks loodud lahendusi.

Kasutatud kirjandus

Anthropic. 2026. *Claude Code (Opus 4.7)* [suur keelemudel]. https://claude.com/product/claude-code. (Vaadatud 17.05.2026).

Barman, Utsab, Amitava Das, Joachim Wagner & Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. Mona Diab, Julia Hirschberg, Pascale Fung & Thamar Solorio (toim), *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 13–23. Doha, Qatar: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-3902.

Cavnar, William B. & John M. Trenkle. 1994. N-gram-based text categorization. *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, 161–175. Las Vegas, USA.

Das, Amitava & Björn Gambäck. 2013. Code-mixing in social media text. *Traitement Automatique des Langues* 54(3). 41–64. https://aclanthology.org/2013.tal-3.3/.

Dunning, Ted. 1994. *Statistical identification of language* (Computing Research Laboratory Technical Memo MCCS 94-273). Las Cruces, NM: New Mexico State University.

Grefenstette, Gregory. 1995. Comparing two language identification schemes. *Proceedings of the 3rd International Conference on Statistical Analysis of Textual Data (JADT 1995)*, 263–268. Rome, Italy.

Jaech, Aaron, George Mulcaire, Shobhit Hathi, Mari Ostendorf & Noah A. Smith. 2016. Hierarchical character-word models for language identification. Lun-Wei Ku, Jane Yung-jen Hsu & Cheng-Te Li (toim), *Proceedings of the Fourth International Workshop on Natural Language Processing for Social Media*, 84–93. Austin, TX: Association for Computational Linguistics. https://doi.org/10.18653/v1/W16-6212.

Jauhiainen, Tommi, Marco Lui, Marcos Zampieri, Timothy Baldwin & Krister Lindén. 2019. Automatic language identification in texts: A survey. *Journal of Artificial Intelligence Research* 65. 675–782. https://doi.org/10.1613/jair.1.11675.

Jauhiainen, Tommi, Heidi Jauhiainen & Krister Lindén. 2022. HeLI-OTS, off-the-shelf language identifier for text. *Proceedings of the Thirteenth Language Resources and Evaluation*

Conference, 3912–3922. Marseille, France: European Language Resources Association. https://aclanthology.org/2022.lrec-1.416/.

Joulin, Armand, Edouard Grave, Piotr Bojanowski & Tomas Mikolov. 2017. Bag of tricks for efficient text classification. Mirella Lapata, Phil Blunsom & Alexander Koller (toim), *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Valencia, Spain: Association for Computational Linguistics. https://aclanthology.org/E17-2068/.

Kargaran, Amir Hossein, Ayyoob Imani, François Yvon & Hinrich Schütze. 2023. GlotLID: Language identification for low-resource languages. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 6155–6218. Singapore: Association for Computational Linguistics. https://aclanthology.org/2023.findings-emnlp.410/.

Kargaran, Amir Hossein, François Yvon & Hinrich Schütze. 2024. MaskLID: Code-switching language identification through iterative masking. Lun-Wei Ku, Andre Martins & Vivek Srikumar (toim), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 459–469. Bangkok, Thailand: Association for Computational Linguistics. https://aclanthology.org/2024.acl-short.43/.

Koreinik, Kadri, Aive Mandel, Maarja-Liisa Pilvik, Kristiina Praakli & Virve-Anneli Vihman. 2022. Outsourcing teenage language: A participatory approach for exploring speech and text messaging. *Linguistics Vanguard* 8(s4). 423–434. https://doi.org/10.1515/lingvan-2021-0116.

Lui, Marco & Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. *Proceedings of the ACL 2012 System Demonstrations*, 25–30. Jeju Island, Korea: Association for Computational Linguistics. https://aclanthology.org/P12-3005/.

Lui, Marco & Timothy Baldwin. 2014. Accurate language identification of Twitter messages. Atefeh Farzindar, Diana Inkpen, Michael Gamon & Meena Nagarajan (toim), *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)*, 17–25. Gothenburg, Sweden: Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1303.

McCandless, Mike. 2013. *Chromium Compact Language Detector 2 (CLD2)* [tarkvarateek]. https://github.com/CLD2Owners/cld2. (Vaadatud 17.05.2026).

Mustonen, Seppo. 1965. Multiple discriminant analysis in linguistic problems. *Statistical Methods in Linguistics* 4. 37–44.

NbAiLab. 2023. *nb-nordic-lid* [keeletuvastusmudel]. Hugging Face. https://huggingface.co/NbAiLab/nb-nordic-lid. (Vaadatud 17.05.2026).

Nguyen, Dong & A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 857–862. Seattle, Washington, USA: Association for Computational Linguistics. https://aclanthology.org/D13-1084/.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot & Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12. 2825–2830. https://www.jmlr.org/papers/v12/pedregosa11a.html.

Reddit Inc. s. a. *User Agreement*. https://redditinc.com/policies/user-agreement. (Vaadatud 17.05.2026).

Stahl, Peter M. 2024. *Lingua: The most accurate natural language detection library for Python* [Pythoni teek]. https://github.com/pemistahl/lingua-py. (Vaadatud 17.05.2026).

Sterner, Igor. 2024. Multilingual identification of English code-switching. Yves Scherrer, Tommi Jauhiainen, Nikola Ljubešić, Marcos Zampieri, Preslav Nakov & Jörg Tiedemann (toim), *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, 163–173. Mexico City, Mexico: Association for Computational Linguistics. https://aclanthology.org/2024.vardial-1.14/.

Verschik, Anna & Helin Kask. 2019. English-Estonian code-copying: Comparing blogs and vlogs. *Applied Linguistics Review* 10(2). Berlin: De Gruyter. https://doi.org/10.1515/applirev-2019-0028.

Vihman, Virve-Anneli, Maarja-Liisa Pilvik, Aive Mandel, Annika Kängsepp, Mari Aigro, Kadri Koreinik, Kristiina Praakli & Liina Lindström. 2023. *Estonian Teen Language Corpus* [andmestik]. Tartu: Tartu Ülikool, eesti ja üldkeeleteaduse instituut. https://doi.org/10.23673/re-455.

Zaragoza-Bernabeu, Jaume. 2024. *heliport: Fast language identifier* [Pythoni teek, Rust-port HeLI-OTS-st]. https://github.com/ZJaume/heliport. (Vaadatud 17.05.2026).

A Comparison of Language Identifiers. Summary.

General-purpose LID reaches near-perfect accuracy on long, well-formed monolingual texts, but its performance drops sharply on the short, non-standard, and code-switched texts typical of social media. Estonian-English code-switching, meaning the use of English words and phrases within Estonian text, is especially common in young people's online and spoken communication. Detection tools suited specifically to the Estonian-English pair, however, have so far received little attention.

The aim of this thesis is to compare twelve widely used language identifiers and assess how well they detect Estonian-English code-switching at the word level. The work pursues three goals:

(1) to compare the identifiers on two corpora of differing character and to derive recommendations for different use cases;

(2) to compile a sufficiently large and varied Reddit-based corpus for analysis;

(3) to prepare the collected material for manual annotation.

The evaluation used two corpora. The first is the Estonian Teen Language Corpus, compiled by Vihman et al. (2023) within the “Teen Speak in Estonia” (TeKE) project. It consists mainly of private Facebook Messenger conversations between Estonian speakers aged 9 to 18 (61,333 tokens, 7.26% of them labelled as English). The second was collected and manually annotated by the author from the r/Eesti subreddit on Reddit (534 sentences, 9,773 tokens, 20.41% of them labelled as English). Both corpora were annotated at the word level following the TeKE annotation scheme, which was minimally adapted to suit the public-forum context.

The twelve identifiers were AnE-LID, lingua (including its multi-language mode lingua-mixed), heliport, GlotLID, nb-nordic-lid, fastText, fast-langdetect, pycld2, masklid (including its code-switching mode masklid-cs), and langid.

Each was assessed using token-level accuracy, precision, recall, F1 score, and a combined overall score. On the chat corpus, heliport achieved the best overall score (74.32%), with AnE-LID coming a close second (71.49%). On the Reddit corpus, AnE-LID achieved the best overall performance (93.49%), clearly outperforming the rest. lingua-mixed came second on the Reddit corpus (87.63%) but only fourth on the chat corpus (64.53%). The identifiers fell into three behavioural groups. The balanced ones (AnE-LID, lingua-mixed, lingua) performed consistently

across precision and recall. The conservative ones (heliport, GlotLID, nb-nordic-lid) reached very high precision on English tokens – heliport scored 87.08% on the chat corpus and 97.43% on the Reddit corpus – but had low recall and left many English tokens undetected. The aggressive ones (masklid, masklid-cs, langid) did the reverse: high recall, low precision, and frequent mislabelling of Estonian words as English.

Every identifier did better on the Reddit corpus than on the chat corpus, which is most likely explained by the Reddit texts being longer and more orthographically standard. For general-purpose word-level detection of Estonian-English code-switching, AnE-LID is the recommended tool, and lingua-mixed is a strong alternative. Where precision matters most, such as automatic corpus enrichment without later manual checking, heliport is the better choice, although it will miss part of the code-switching. fastText and fast-langdetect are adequate when the balance between precision and recall is not critical. langid, masklid, and masklid-cs are not recommended for this task, since their overall scores stayed below 65% on both corpora.

Word-level detection of Estonian-English code-switching is therefore a feasible task, but only with a language identifier that was designed for multilingual text, or at least adapted to it. General-purpose sentence-level identifiers, even technically more advanced ones, cannot replace tools built for code-switching.

Keywords: code-switching, language identification, corpus linguistics, social media

Lisa 1. Märksõnade loetelu Redditi andmete kogumisel

Üldised eesti sõnad: minu, meie, tema, nende, kõik, palju, väga, hästi, arvan, mõtlen, tegelikult, ilmselt, vist, kindlasti, probleem, küsimus, vastus, teema, asi, elu, inimesed

Küsimused ja arutelud: kuidas, miks, kas keegi, kas on, mis arvate, soovitan, kogemus, nõuanne, abi, help, advice, recommendation

Tehnoloogia: IT, startup, programmeerija, coding, tech, developer, arvuti, software, app, AI, crypto, blockchain, telefon, internet, online, website, server, database

Meelelahutus: Netflix, film, muusika, gaming, YouTube, Spotify, seriaal, mängud, anime, podcast, TikTok, stream, movie, show, series, concert, festival

Igapäevaelu: töö, kool, ülikool, toidukott, shopping, rent, korter, palk, töökoht, job, interview, remote, elu, päev, hommik, õhtu, nädalavahetus, puhkus

Emotsioonid ja arvamused: viha, rõõm, kurb, ilus, naljakas, imelik, hull, parim, halvim, lemmik, vihkan, armastan, meeldib, cringe, weird, funny, crazy, best, worst

Suhted: sõber, pere, vanemad, lapsed, abikaasa, kallim, suhe, dating, tinder, relationship, breakup, wedding

Raha ja finantsid: raha, palk, hind, odav, kallis, soodne, ost, pank, laen, krediit, investeering, säästmine, salary, price, expensive, cheap, budget, loan

Sport ja fitness: jalgpall, NBA, fitness, maraton, gym, training, jooks, sport, treening, võistlus, UEFA, workout

Släng ja vaba kõnekeel: lmao, btw, tbh, ngl, cringe, vibe, mood, literally, random, nice, cool, based, sus, basically, actually, honestly, seriously

Toit ja elustiil: restoran, kohvik, recipe, vegan, burger, pizza, õlu, cocktail, brunch, delivery, Wolt, Bolt, söök, toit, retsept, küpsetamine, cooking

Reisimine: reisimine, travel, Tallinn, lennuk, flight, vacation, hotel, airbnb, tourist, viisa, passport, trip

Tervis ja vaimne tervis: tervis, arst, haigla, diagnoos, ravi, ravim, vaimne, depressioon, ärevus, stress, burnout, doctor, health, therapy, mental, anxiety

Eluase ja elamispind: korter, maja, üür, ost, müük, remont, sisustus, naaber, ühistu, elamispind, rent, apartment, house

Haridus: õppimine, eksam, kraad, bakalaureuse, magistri, kursus, loeng, professor, study, degree, university

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Marko Sasi,

1) annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Keeletuvastajate võrdlus", mille juhendaja on Kadri Muischnek, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

2) annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi kaudu Creative Commonsi litsentsiga CC BY NC ND 4.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni;

3) olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile;

4) kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

/allkirjastatud digitaalselt/

21.05.2026