TARTU ÜLIKOOL

Matemaatika-informaatikateaduskond

Arvutiteaduse instituut

Informaatika eriala

Hendrik Aruoja

Automaatmärgendatud süntaksi ja morfoloogia konfliktide lahendamine

Bakalaureusetöö (9 EAP)

Juhendaja: Sven Laur DSc. (Tech)

Tartu 2025

Abstract

Resolving Conflicts in Automatically Annotated Syntax and Morphology

The Bachelor's thesis focuses on finding cost-effective ways to resolve syntax-morphology conflicts for EstNLTK that utilizes instructing of Large Language Models (LLMs) as an alternative to manually labeling. The goal is to create a Python scripts that includes database query generation, execution of queries and processing of LLM prompts via API. The work is divided into a theoretical part, where the principles of syntax and morphology analysis are introduced, and a practical part, where the created python scripts and their capabilities are described.

Keywords: Python programming language, Natural Language Processing, Large Language Models, Syntax and morphology

CERCS: P176 Artificial intelligence

Kokkuvõte

Automaatmärgendatud süntaksi ja morfoloogia konfliktide lahendamine

Bakalaureusetöö keskendub kuluefektiivse süntaksi-morfoloogia konfliktide märgendamise viisi leidmiseks EstNLTK jaoks, kasutades suurte keelemudelite (LLM-idel) instrueerimisel põhinevat lähenemist käsitsi märgendamise alternatiivina. Eesmärgiks on luua Pythoni skriptid, mis sisaldavad andmebaasipäringute genereerimist, nende käivitamist ning instrueeritavatele keelemudelitele saadetavate päringute töötlemist API kaudu. Töö koosneb teoreetilisest osast, kus tutvustatakse süntaksi ja morfoloogia analüüsi põhimõtteid, ning praktilisest osast, kus kirjeldatakse loodud skriptide ülesehitust ja toimimist.

Märksõnad: Programmeerimiskeel Python, loomuliku keele töötlus, suured keelemudelid, süntaks ja morfoloogia

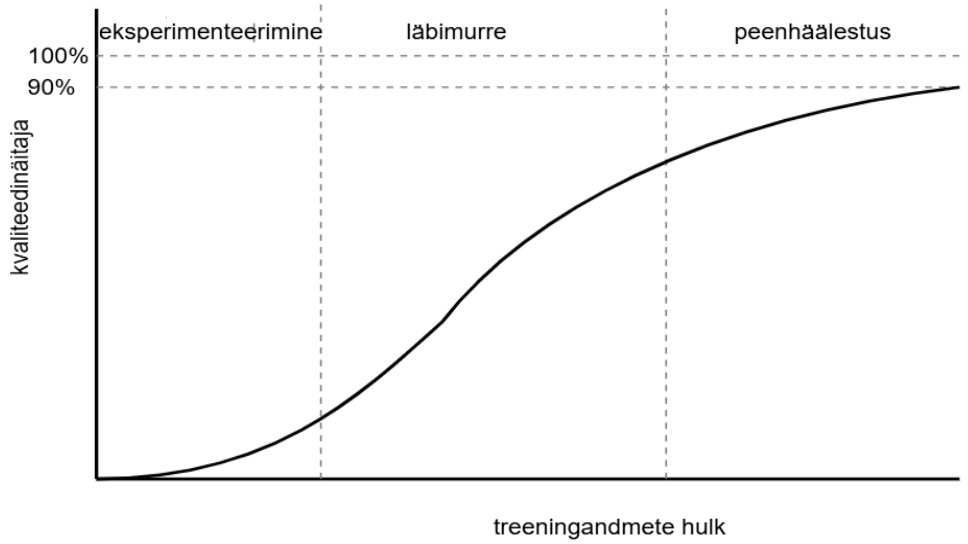
CERCS: P176 Tehisintellekt

Sisukord

Sissejuhatus

Moodsate automaatsete morfoloogilise ja süntaksianalüüsi meetodite kasutamine eeldab kvaliteetselt ja täpselt märgendatud näidisandmete olemasolu. See protsess on sageli ajaliselt ja rahaliselt kulukas, kuna iga analüüsitud teksti osa (nt sõnad, fraasid ja laused) tuleb seostada õige morfoloogilise ja süntaktilise kategooriaga. Täpsed märgendamisandmed on hädavajalikud mudelite täpsuse tagamiseks, kuid nende kogumine on töömahukas ja aeganõudev. Kõrge kvaliteedi tagamiseks tuleb kasutada kogenud lingvisti abiga loodud andmestikke, mis on sageli piiratud ulatusega ja ei kata kõiki võimalikke keelekasutuse nüansse.

Üks suurimaid probleeme automaatsete analüsaatoritee treenimisel on S-kurv (kõver, mis kirjeldab mudeli kvaliteedi paranemist treenimise käigus). S-kurv ilmestab olukorda, kus juba kõrge õigsuse (nt 90–95%) saavutanud mudeli treenimine täiendava täpsuse saavutamiseks muutub üha keerukamaks ja aeglasemaks. See tähendab, et kui läheneme lõppeesmärgile (100% täpsusele), siis muutub täiendava progressini jõudmine ja sobivate treeningandmete leidmine järjest keerulisemaks. Ilma nutika andmevalikuta võib mudelite täiendamiseks olla vaja tohutult suuri andmekogumeid, mis omakorda suurendab arendus- ja hoolduskulusid (*vt joonis 1*).



Joonis 1. S-kurv näitab, kuidas täpsuse lähenedes 100%-le hakkab vajalike treeningandmete hulk järjest kiiremini suurenema. Eksperimenteerimise faas on tüüpiliselt baasuuringute alguses, läbimurre

tähendab tüüpiliselt kommertsialiseerimist või praktikas rakendamist ning peenhäälestuse faasi jõutakse aastatepikkuses süsteemide ja firmade omavahelises võistluses.

Kuna täielikult märgendatud andmete kogumine on töömahukas ja ressursirohke, on üks võimalik lahendus analüüsida erinevate lingvistiliste märgenduste vahelist kooskõla. Selle töö raames räägime morfoloogia ja süntaksi kooskõlast. Morfoloogia analüüsib sõnade struktuuri, tuletades sõnade algvorme ja grammatilisi omadusi, samas kui süntaks analüüsib sõnade omavahelist seost lauses. Kooskõla nende kahe analüüsi vahel on oluline, kuna süntaktiliste vigade parandamine võib samal ajal aidata tuvastada morfoloogilisi vigu.

Morfoloogia ja süntaksi kooskõla võib avalduda erineval kujul, üks lihtsamaid ja sagedasemaid näiteid selleks on rektsioon.

Rektsioonid kirjeldavad, kuidas erinevad sõnad (nt tegusõnad ja nimisõnad) peaksid üksteisega kokku sobima vastavalt keele grammatika reeglitele. Näiteks võib verb „võitlema" korral võib eeldada, et sellele järgneb nimisõna, mis esindab objekti või sihtmärki, nagu „võitlema vaenlasega". Kui keeleline rektsioon on valesti määratud, siis võib see põhjustada süntaktilisi vigu või segadust analüüsitavas tekstis [1].

Käesolevas töös uuritakse verbi ja nimisõnafraaside rektsioone, keskendudes sellele, kuidas neid rektsioone saab rakendada morfoloogia ja süntaksi vaheliste mittekooskõlade analüüsimisel. Kuigi lingvistid on kureeritult andmeid rektsioonide kohta kogunud, ei ole need andmed täielikud ja ei kata kõiki võimalikke rektsioonide kombinatsioone, mis tegelikus keelekasutuses esineda võivad.

Töö aluseks olevaks andmestikus on kokku kogutud näited vastuoluga morfoloogilise ja süntaktilise analüüsi vahel. Need vastuolud ilmnevad siis, kui sõnade morfoloogiline analüüs ei sobitu korrektselt lause struktuuriga või kui süntaks ei arvesta õigete morfoloogiliste tunnustega. Ülesanne on tuvastada, milles seisneb probleem: kas see on tingitud valest morfoloogilisest analüüsist, vigasest süntaksipuust või puudulikust rektsiooni kirjeldusest.

Töö eesmärk on vähendada oluliselt andmestiku edasiseks kureerimiseks vajalikku inimtöö hulka. Selleks on vaja välja töötada meetodeid, mis suudavad tuvastada ja parandada keelelisi vigu tõhusalt, kasutades olemasolevaid instrueeritavaid keelemudeleid (IKM) ja analüüsitehnikaid. Täiendav eesmärk on automatiseerida rektsioonide määramise protsess, mis võib vähendada inimeste tööjõu vajadust ja suurendada analüüsi täpsust.

Teoreetiline taust

Lause süntaktiline struktuur määrab eesti keeles selles osalevate sõnade morfoloogilise kuju. Morfoloogia ja süntaksi konflikt tekib siis, kui sõnade grammatilised vormid (morfoloogia) ja nende lausesisesed suhted (süntaks) ei ühti või viitavad erinevale tõlgendusele. See juhtub eriti sageli eesti keeles, kus sõnade käändevormid võivad olla mitmetähenduslikud – sama vorm võib täita erinevaid süntaktilisi rolle.

Morfoloogiline analüüs

Morfoloogiline analüüs on keeleteaduses ja loomuliku keele töötluses (NLP) protsess, mille käigus uuritakse sõna kuju ja sisemisi struktuure, tuvastades selle morfeemid, tüvi, lemma, tunnused, näiteks käänded ja pöörded, ning võimalikud morfoloogilised variandid. Arvutuslingvistikas on see oluline, et määrata sõna morfoloogiline struktuur.

Lemma, vorm ja sõnaliik on kolm olulist keeleanalüüsi mõistet. **Lemma** on sõna põhikuju, see tähendab kuju, mille järgi sõna leidub sõnaraamatus – näiteks sõnast „majades“ on lemma „maja“ ning sõnast „läksid“ on lemma „minema“. **Vorm** kirjeldab sõna konkreetset grammatilist kuju lauses, nagu arv, kääne, aeg, või pöördevorm. Näiteks „majades“ on mitmuse seesütlev ja „läksid“ on lihtmineviku, 2. pöörde ainsus. **Sõnaliik** näitab, millisesse grammatilisse kategooriasse sõna kuulub – näiteks nimisõna, tegusõna, omadussõna või asesõna. Nii on „maja“ nimisõna (S), „lugema“ tegusõna (V) ja „ilus“ omadussõna (A). Need kolm tasandit – lemma, vorm ja sõnaliik – aitavad mõista, kuidas sõna toimib keeles nii oma tähenduse kui ka grammatika kaudu.

Morfoloogiline annotatsioon on protsess, mille käigus sõnadele lisatakse märgendid, mis näitavad nende grammatilist kuju, sõnaliiki ja põhivormi (lemmat), et võimaldada nende automaatset keelelist analüüsi. Morfoloogilist analüüsi alustatakse sageli lemmatiseerimisest, sest see aitab vähendada mitmetähenduslikkust. Lemmatiseerimisest järgmine samm on märgendite lisamine. Märgendite lisamiseks on eesti keele jaoks loodud töövahend Vabamorf [2], mis suudab teha morfoloogilist analüüsi, tuvastada sõnade põhivormid, sõnaliigid ning mitmesugused grammatilised tunnused, nagu kääne, arv, pööre ja tegumood.

Antud töös on Vabamorfi kasutatud ainult EstNLTK[3] vahendusel. See tähendab, et morfoloogiline analüüs toimus läbi EstNLTK liidese, ilma et oleks Vabamorfi eraldi käsurealt või otse koodi kaudu välja kutsutud. Vabamorf ja EstNLTK kasutavad märgendamiseks EstMorf nimelist märgendussüsteemi (vt Tabel 1). Täielik ülevaade Vabamorfi morfoloogiliste kategooriate kohta on EstNLTK dokumentatsioonis [4].

Sõna	Lemma	Sõnaliik	Vabamorf EstMorf
Poiss	poiss	S (nimisõna)	poiss S sg n (ainsus, nimetav)
lõi	lööma	V (tegusõna)	lööma V af s3 ps afm (minevik, ainsus, 3. pööre)
lõi	looma	V (tegusõna)	looma V af s3 ps afm (minevik, ainsus, 3. pööre)
valge	valge	A (omadussõna)	valge A sg n (ainsus, nimetav, seotud nimisõnaga „palli“)
palli	pall	S (nimisõna)	pall S sg n adt (ainsus, osastav)
jalaga	jalg	S (nimisõna)	jalg S sg com ad (ainsus, kaasaütlev)
vastu	vastu	K (kaassõna)	vastu K (kaassõna)
seina	sein	S (nimisõna)	sein S sg n adt (ainsus, osastav)
.	.	Z (kirjavahemärk)	

Tabel 1. Näide ühest Vabamorf poolt märgendatud lausest.

Kasutades Vabamorf analüsaatorit saab sõnade morfoloogilist analüüsi teostada mitmel viisil. Oletusel põhinev analüüs tugineb tõenäosusele, et teatud morfoloogilised struktuurid esinevad rohkem kui teised. See võimaldab analüsaatoril esitada rohkem kui ühe võimaliku tõlgenduse ning tulla toime ka sõnastikuväliste sõnadega (nt uuemad tuletised, eksitused, nimed jne). Reeglipõhine analüüs kasutab fikseeritud morfoloogilisi reegleid, et määrata sõna morfoloogiline struktuur.

Disaini poolest annab Vabamorf analüsaator välja analüüse ilma ühestamiseta, seega on selle väljundikiht ambivalentne (mitmetimõistetav). Ligikaudu 45% eesti keele sõnavormidest tekstides omab rohkem kui ühte kehtivat analüüsi [5].

Lisaks saab kasutada ka heuristilist ühestamist. Ühestamise eesmärk on valida morfoloogilise analüüsi võimalikest tulemustest kõige tõenäolisem. Vabamorfis kasutatav ühestaja on hetkel HMM-põhine (Markovi peitmudel), mis kasutab statistilist tõenäosust, et otsustada, milline morfoloogiline analüüs on konteksti põhjal kõige sobivam. Kuna Vabamorf ühestaja on lihtne ja väikesemahuline tööriist, mis põhineb piiratud kontekstil, ei kasuta semantikat, siis võib seetõttu keerulisematel juhtudel eksida. [5].

Vabamorfi nimisõna käände määramise täpsuse hinnang sõltub sellest kas mitmetähenduslikel juhtudel valitakse juhuslik või esimene analüüs kandidaatide nimekirjast. Need meetodid annavad üldiseks täpsuseks vastavalt 89,76% ja 90,29%. Kvaliteetse morfoloogilise annotatsiooniga Eesti puudepanga andmestikul (Estonian Dependency Treebank) on analüüside täpsuseks 83,61% korpuse sõnadest [6].

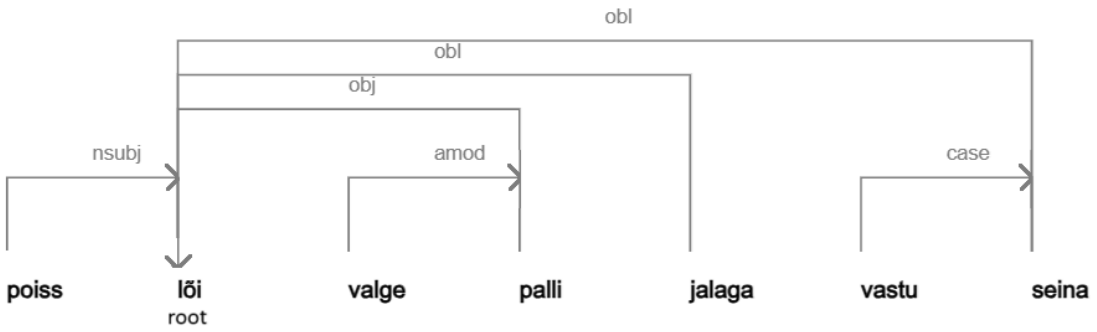
Süntaksianalüüs

Süntaktiline analüüs tuvastab lause sisemise struktuuri: leiab fraasid ja nendevahelised sõltuvused. Süntaksianalüüsi väljundiks on on kas fraasissüntaks ja sõltuvussüntaks. Süntaksianalüüsi väljundiks on tavaliselt süntaksipuu (*syntax tree* või *parse tree*), mis on graafiline esitus lause süntaktilisest struktuurist, mis näitab, kuidas lause koosneb erinevatest osadest ja kuidas need osad omavahel grammatilistes suhetes on.

Fraasisüntaks (*constituency grammar*) kirjeldab lause ülesehitust hierarhiliste fraaside kaudu, nagu nimisõnafraas ja tegusõnafraas, mis omavahel kombineeruvad suuremateks üksusteks. See lähenemine näitab, kuidas sõnad rühmituvad ja milliseid struktuurilisi rolle nad fraasis täidavad [7].
Praegu ei ole fraasisüntaksi kasutamine eesti keeles laialdaselt levinud ning vajalikud töövahendid eesti keele töötlemiseks puudulikud.

Sõltuvussüntaks (*dependency parsing*) keskendub sellele, kuidas iga sõna lauses on seotud teistega otseste sõltuvussuhete kaudu nagu subjekt või sihitis. Selle lähenemise eesmärk on pakkuda ühtset ja lihtsat analüüsi, mis sobib automaatseks keele töötlemiseks eri keeltes. Sõltuvussuhted kujutatakse tavaliselt suunatud puudena, kus iga sõna on seotud mõne teise sõnaga alluvussuhtes, välja arvatud lause peasõna ehk juur, mis ei allu kellelegi [7].
Kaar näitab alluvussuhet - millise sõna tähendust alluv täpsustab. Näiteks lauses *poiss lõi palli* on juureks sõna lõi (lööma) ja alluvussuhted näidatud nooltega poiss → lõi ja palli → lõi.
Sõltuvussüntaksipuus on igal sõnadevahelisel kaarel ka märgend, mis täpsustab alluvussuhet (vt joonis 2).

UD-süntaks (Universal Dependencies Syntax) on teadlaste poolt arendatav sõltuvussüntaks, mis on mõeldud paljude keelte analüüsiks. Sõltuvussuhte märgendid ja nende tähendus on fikseeritud universaalselt üle keelte. UD-süntaks on arendatud masinmärgendamise otstarbel ja disainitud sellisena, et masinmärgendamine oleks lihtsam [8].



Joonis 2. Sõltuvussüntaksipuu lausele “poiss lõi valge palli jalaga vastu seina”.

Stanza on masinõppepõhine loomuliku keele töötluse (NLP) tööriistakomplekt, mille on välja töötanud Stanfordi Ülikool. See võimaldab automaatselt analüüsida lauseid erinevates keeltes, kasutades UD-süntaksiskeemi, ja on treenitud ka eestikeelsel andmestikul. Viimane Stanza versioon on mitmekeelne *end-to-end* süsteem, mis võimaldab süntaksi- ja morfoloogiaanalüüsi teha ühtse protsessina mitmes keeles. EstNLTK kasutab süntaksi märgendamiseks Stanza mudelit, mis saab lisasisendina rikastatud morfoloogilise analüüsi kihi, kuna see on täpsem kui stanza poolt pakutav end-to-end mudel [9].

Nimisõna-verbi rektsioonid

Süntaktilise struktuuri ja morfoloogilise vormi kooskõla kirjeldavad paljud reeglid. Tüüpiliseks viisiks selliseid reegleid esitada on rektsioonid. Üldisemas mõttes sõltumus ehk rektsioon on selline sõnadevaheline seos, kus põhjaks oleva sõna tähendus määrab tema laiendiks oleva sõna käände- või pöördevormi või temaga seotud kaassõna. Selles töös kasutusel olev nimisõna-verbi rektsioon on keeleline suhe, mis määrab, kuidas verb või nimisõna ühendatakse teise sõnaga (näiteks nimisõnaga).

Verbide argumentstruktuur ja rektsioonimallid määravad, millistes käändevormides esinevad verbi poolt nõutavad lause liikmed. Tavaliselt määravad verbid kindla hulga verbimalle, millest igaüks täpsustab verbiga seotud fraaside süntaktilisest märgendusest täpsema klassifikatsiooni. Samuti on nimisõnadel semantilised tähenduspiirangud, mis määravad, millistesse seostesse nad saavad verbidega astuda. Näiteks lauses *Lõin jalaga palli vastu seina* tähistavad „jalaga“ instrumenti, „palli“ sihitist ja „vastu seina“ sihtkohta. See näitab, kuidas süntaktilised suhted on seotud sõnade morfoloogilise vormiga ning aitavad keeleanalüüsis mõista keelelisi seoseid [10].

Rektsiooni illustreerimiseks võib võtta sõna „lööma/lüüa“, mille rektsioon sõltub kontekstist: *keda? mida? – ära löö mind!*; *kuhu? – Valu lõi südamesse*; *millega? milleks? – Mees lõi rusikaga vastu lauda*, *Mees lõi kepi tükkideks*. Võimalik on ka rektsioon *kelleks? – Mees löödi rüütliks*.

EKI on kogunud rektsioone ning masinloetaval kujul on neist väike osa esitatud Tabelis 2, mis sisaldab käsitsi koostatud rektsioone (ainult 3325 rida) ja on väiksem, kuna rektsioonisõnastikust on välja jäetud keeleliselt triviaalsed rektsioonid. Tabeliga tutvudes on samuti ilmne, et see ei sisalda kõiki keeles esinevaid rektsioone – näiteks sisaldab see verbiü lõi palli", st "lööma mida".

wordid	lexeme_id	word	homonym_nr	is_public	pos	government
189803	1226838	käega lööma	1	True	v	millele
190985	1229837	külge lööma	1	True	v	kellele
198003	1247282	läbi lööma	1	True	v	kus
198003	1247283	läbi lööma	1	True	v	mida

198003	1247287	läbi lööma	1	True	v	mida
198519	1248725	lööma	1	True	v	kuhu

Tabel 2. Väljavõte EKI Ühendsõnastikus olevast rektsioonide tabelist 21.08.24 seisuga [11].

Selleks, et seda mittetäielikku rektsioonide kogu täiendada automaatselt tuvastatud seostega, puuduvad kahjuks eesti keele jaoks nii piisavalt ulatuslikud verbimallide andmebaasid kui ka sõnade semantiliste omaduste andmebaasid.

On hulgaliselt rektsioonimalle, mis jäävad samaks paljude verbide korral. Näiteks subjekt ehk alus väljendatakse eesti keeles reeglina nimetavas käändes (*poiss lõi palli, poiss ütles midagi*). Kuna aluse kääne ei sõltu verbist või lause muudest elementidest, siis on subjekti rektsioon tabelist välja jäetud. Samas teised rektsioonimallid sõltuvad konkreetsest verbidest. Näiteks sihitis ehk objekt võib olla väljendatud erinevates käänetes. Näiteks *Peeter sööb suppi*, *ma leidsin suure kärbseseene*, *uks avati*. [12]

Ressursid

Töös kasutatavateks põhilisteks ressurssideks on eesti keele koondkorpuse verbifraaside andmebaas ja instrueeritavad keelemudelid. Verbifraaside andmebaas sisaldab lauseid ja verbide seosed ülejäänud lauseosadega.

Koondkorpus

Eesti keele koondkorpuses on terviktekstid, mitte 2000-sõnalised tekstikatked, millest suures osas koosneb korpusest Eesti Kirjakeele Korpus 1890-1990. Korpus sisaldab ainult kirjalikku keelekasutust ja sisaldab terviktekste erinevatest allikatest, nagu ilukirjandus, ajalehed, teadusartiklid ja seadused. Korpus pakub väärtuslikku materjali keele analüüsiks ja teadusuuringuteks, sisaldades miljoneid sõnu. Kõik tekstid on märgendatud TEI failiformaadi spetsifikatsiooni järgi ja sisaldavad teavet algvormide, grammatiliste kategooriate ja süntaktiliste funktsioonide kohta. Korpus on vaba kasutamiseks mitteärilistel eesmärkidel [13].

Käesolevas töös on koondkorpuste all mõeldud vaid lausete kogumit ja mitte selle analüüse, sest need on EstNLTKga uuesti üle märgendatud. Lisatud on parem ja ühtlasem sõnastus, morfoloogia ja süntaks.

Koondkorpuse verbifraaside andmebaas

Koondkorpuse verbifraaside andmebaas on Katrin Tsepelina ja Sven Lauri teadusprojekti raames koostatud ressurss [14], mis keskendub eesti keele verbifraaside kogumisele ja analüüsimisele. Andmebaas sisaldab lauseid, kus verbid on seotud erinevate sõltuvustega, andes põhjaliku ülevaate verbide ja nende rektsioonide mitmekesisusest eesti keeles. Selle eesmärk on toetada keeleteadusealaseid uuringuid, pakkuda väärtuslikke andmeid masinõppe mudelite treenimiseks ning aidata arendada keele analüüsi tööriistu, nagu morfoloogilised analüsaatorid ja süntaksiparsijad.

Kõigi verbide esinemiste tabel (*transaction_head*) sisaldab kõiki verbide esinemisi koos nende asukoha, vormi ja grammatiliste tunnustega (vt Tabel 3). Lisaks on ära toodud ka lihtsustatud fraas – verb koos kõigi selle otseste alluvatega. Verbi alluvate tabel (*transaction_row*) kirjeldab iga eelmises tabelis oleva verbi otseseid alluvaid süntaksipuus.

Iga rida kirjeldab ühte otses alluvat koos vastavate morfosüntaktiliste tunnustega, nagu *feats*, *deprel*, ja kogu fraasi tekst. Verbide salvestamisel võeti tabelisse vaid sellised verbid, mille verbi aeg oli üks järgnevatest: minevik, imperfekt, olevik ning mille verb polnud umbisikulines vormis (vt Tabel 4).

veerg	kirjeldus	näide
id	t	3
sentence_id	lause indeks	5
loc	positsioon	11
verb	tegusõna	saama
verb_compound	verbitäiend	pihta
form	vorm	saanud
deprel	sõltuvussuhe	root
feats	lingvistilised tunnused	aux,partic,past,ps
phrase	fraas	Bändi kidramees ei saanud keeltele pihta

Tabel 3 . Tabelisse *transaction_head* on kogutud kõigi verbide esinemised korpuses.

veerg	kirjeldus	näide
id	indeks	4
head_id	peasõna indeks	3
loc	asukoht fraasis	1
loc_rel	suhteline asukoht fraasis	-3
deprel	sõltuvussuhe	obj
form	vorm	Bändi
lemma	algvorm	bänd
feats	lingvistilised tunnused	adit,com,sg
parent_loc	vanema asukoht	NULL
pos	lauseosa (part of speech)	S

Tabel 4. Tabelisse *transaction_row* on kogutud tabelis *transaction_head* oleva verbiga otseses alluvuses olevad sõnad.

Süntaksi ja morfoloogia konfliktide andmebaas

Algandmetena on kasutada Katrin Tsepelina ja Sven Lauri koostatud süntaksi ja morfoloogia konfliktide andmebaas [15]. See andmebaas loodi eesmärgiga tuvastada olemasolevate mudelite vigu ja parandada nende täpsust. Sellesse tabelisse on kogutud juhud, kus rektsioonipõhine ja oletuspõhine analüüs on jõudnud erinevate tulemuseni.

Tabelis on olemas kõik morfoloogiliseks analüüsiks vajalikud tunnused, mis kirjeldavad lause struktuuri ja grammatikat. Iga rea kohta on esitatud erinevad üksikasjad, näiteks rektsioonimalli indeks, tegusõna (punane), tegusõna asukoht ja algse llause indeks (hall), mis võimaldab vajadusel täpsemalt uurida, millisest täispikast lausest korpuses ta pärineb. Sinisega on toodud huvipakkuv nimisõna. Rektsioonipõhise analüüsi andmeväljad on esitatud rohelisega ja oletuspõhise analüüsi tulemused on välja toodud oranžiga (vt Tabel 5).

id	rea indeks	1
pattern_id	verbimalli indeks	1
sentence_id	algse lause indeks	30442
verb_loc	tegusõna asukoht	3
compound_loc	verbitäiendi asukoht	null
phrase_root_loc	uuritava nimisõna asukoht	4
verb_phrase_loc	fraasi sõnade positsioonid algses lauses	[1, 2, 3, 4]
phrase_case	kääne rektsiooni järgi	part
phrase_deprel	sõltuvussuhe	obl
verb	tegusõna	aasima
verb_compound	tegusõna täiend	-
phrase	uuritav fraas	Ma siis aasisin **Deani**

phrase_root_lemma	uuritava fraasi algvorm	Dea
current_analysis oranz	analüüs	prop,sg,term
current_case oranz	ühestatud kääne	term
possible_cases	võimalikud käänded	[null, nom, term, part, gen, adit]

Tabel 5. Ühe tabeli rea veergude tähendused

Instrueeritavad keelemudelid

Käesolevas bakalaureusetöös on kasutatud instrueeritavate keelemudelite abi, et mitte käsitsi märgendada. Käsitsi märgendus on liiga kulukas – näiteks 1000 sõna märgendamine võib maksta umbes 100 eurot. Samal ajal suudavad instrueeritavad keelemudelid sünteesida keeleliselt korrektset teksti, mis viitab sellele, et nad peavad konteksti põhjal mõistma ka seda, mis käändes mingi sõna on või saab olla.

Instrueeritavad keelemudelid on suured keelemudelid (*Large Language Models, LLMs*), mida on täiendavalt treenitud kasutaja instruktsioonidele võimalikult hästi vastama. Instrueeritavatele keelemudelitele on võimalik anda instruktsioon (*prompt*) ja sisend (*input*), mille peale mudel annab väljundi. Tänapäeval on valdav enamik instrueeritavaid keelemudeleid ehitatud transformeritel põhinevale arhitektuurile (vt Tabel 6).

Mudel	Arendaja	Kättesaadavus kinniste kaaludega (pilvepõhine, läbi API) / avatud kaaludega (lokaalselt kasutatav)
ChatGPT-4o	**OpenAI**	**kinnine**
T5	Google	avatud
Claude	Anthropic	kinni
Llama 3.1	**Meta (Facebook)**	**avatud**
DeepSeek	DeepSeek	avatud

Tabel 6. Tuntumad transformer tüüpi mudelid, sh antud töös kasutuses olevad ChatGPT-4o ja Llama 3.1.

Antud töö kontekstis huvitab meid ainult nimisõna käändevorm. Instrueeritavad keelemudelid saavad suurepäraselt aru käändevormidest, kuna nad suudavad genereerida korrektset eestikeelset teksti. Kuigi neid pole spetsiaalselt treenitud grammatika tundmiseks, omandavad nad keelelised struktuurid implitsiitselt. Samas instrueeritavad keelemudelid pole loodud süntaktiliseks ega morfoloogiliseks analüüsiks ja neilt otse käände küsimine pole otstarbekas, seega tuleb leida uus lähenemisviis.

Üks võimalus on asendada fraasis huvipakkuv sõna mõne ühestamist mittevajava sõnaga. Need esinevad tavaliselt unikaalsetes käänetes ja nende käänet suudab töökindlalt määrata näiteks Vabamorf. Nii saab keelemudeli võimekust kasutada kaudselt – asendussõna kaudu järeldades algse sõna käände.

Antud töös on oluline ka instrueeritavate keelemudelite hinnastamine, et arvutada sellisel viisil märgendamise kulu. Näiteks OpenAI arvestab hinda tokenite alusel, kus **1 token vastab ligikaudu 0,75 sõnale** (vt Tabel 7). Tehtud katsetes kujunes ühe märgenduse hinnaks suurusjärk 0,00075 eurot.

Mudel	Sisend (1K tokenit)	Väljund (1K tokenit)	Kokku (USD / 1K tokenit)	Ligikaudne hind ühe lause kohta (15 tokenit)
GPT-3.5 Turbo	$0.0005	$0.0015	$0.002	$0.00003
GPT-4 Turbo	$0.01	$0.03	$0.04	$0.0006
GPT-4	$0.03	$0.06	$0.09	$0.00135
GPT-4 32k	$0.06	$0.12	$0.18	$0.0027
GPT-4o	**$0.005**	**$0.015**	**$0.02**	**$0.0003**
GPT-4o mini	$0.00015	$0.0006	$0.00075	$0.00001125

Tabel 7. Siin töös kasutatava mudeli ChatGPT-4o hind võrrelduna teiste mudelitega.

Arvestades, et ChatGPT-4o puhul tuli päringute koguarv suurusjärgus 20 000 ja keskmine päringu pikkus oli suurusjärgus 30 sõna läks kogu analüüs ChatGPT-4o APIt kasutades maksma ligikaudu 20 eurot. Lokaalse mudelina oli kasutuses töö alustamise hetkel parim saadaolev, avatud kaaludega 405 miljardit parameetriga Llama 3.1, mis töötas Tartu Ülikooli High Performance Computing keskuses (teenusel Rocket/pegasus). Arvestades, et selle teenuse hinnastamine on tunnihinna arvestuse baasil (mälu 0.012 EUR/6GB/h, GPU 0.5 EUR/GPU/h) ,

kasutuses oli 64GB mälu ja server oli kokku kasutuses ligikaudu 24h, saame kogumaksumuseks samuti suurusjärgu 27 eurot. [16]. Seega oli mõlema mudeli kasutamise maksumus ligikaudu samas suurusjärgus ja kuigi hetkel saadaolevate versioonide võrdluses näitas ChatGPT-4o kohati paremaid tulemusi kui Llama 3.1 siis sisuliselt ei olnud suurt vahet kumba mudelit kasutati.

Andmete analüüs

Konfliktide tabelis on 655 000 juhtumit, mille korral rektsioonipõhine ja oletusel põhinev analüüs annavad erineva tulemuse. Need juhtumid jagunevad kuueks alamhulgaks. Ja iga alamhulk jaguneb omakorda kolmeks (IKM arvates korrektne oletuspõhine analüüs, IKM arvates korrektne rektsioonipõhine analüüs ja juhud, kus IKM hinnang ei lange kummagiga kokku (vt Tabel 8).

Alam hulk	Näitefraas	Oletuslik kääne	Rektsioonipõhine kääne
I	**Jääkainetest** aitab vabaneda	nom	el
II	ahistanud **end** programmiga	-	part
III	**Tallinnaga** aitavad pidada marsruuttaksod	kom	part
IV	kindlus on **Andyt** aidanud	nom	part
V	Kõik ahvivad praegu **saundi**	nom	part
VI	Primakov adresseeris **selle** aga Rjurikovile	gen	all

Tabel 8. Tüüpilised näited fraasidest, milles analüüsitava sõna kääne tuvastati erinevalt reeglipõhise ja oletusliku mudeli poolt. Alla on joonitud õigesti tuvastatud kääne.

Konfliktide alamhulkadeks jagamine

Mõistlik on vastuoluliste analüüside andmebaas jagada väiksemateks alamhulkadeks. Sellisel moel eraldades jäid alles teatavate omadustega konfliktide alamhulgad. Konfliktsete analüüside kirjeid uurides ilmnesid teatavad mustrid, mis võimaldavad teha mõistlikke oletusi konflikti tekkimise põhjuste kohta. Jagamise kriteeriumiks olid hüpoteesid, miks konflikt tekib. Algandmestik jaguneb kuueks suuremaks alamhulgaks, vastavalt fraasi struktuurile ja sõna omadustele. Esimene alamhulk sisaldab juhtumeid, kus kääne oli üheselt määratav; teises alamhulgas olid peamiselt määrsõnad (käändumatud sõnad), mille hulgas üksikud nimisõnad. Kolmas alamhulk hõlmas juhtumeid, kus analüüsitav sõna oli fraasi alguses. Neljandas rühmas olid pärisnimed, mille puhul Vabamorfil on raskusi nii algvormi kui ka käände määramisel. Viies alamhulk sisaldab nimisõnasubjekte, mille puhul võimalike käänete hulgas oli osastav. Näiteks *partitsioone aitab teha.* Kõik muud juhtumid liigitati kuuendasse, määratlemata alamhulka (vt Tabel 9).

Alam hulk	Kirjeldus ja näide	Ridu
I	**Ainuvõimalik kääne.** Selles alamhulgas on rektsioonil põhineval analüüsil vaid üks ainuvõimalik kääne.	17322
	fraas: **Sportlasi** arektsiooni võimalikud käänded: [**part**]histas seegi rektsioonil põhinev kääne: **part** oletuslik kääne: nom	
II	**Määrsõna**. Võimalike analüüside arv on 2 ja teine neist <null>. Valdavas osas sellistel juhtudel analüüsitava sõna liik (POS) ei ole nimisõna. Selles alamhulgas on sõnad, mis vaatamata identse morfoloogilise kuju poolest võivad olla nii määrsõnad kui ka mingis käändes nimisõnad. Sellest alamhulgast on näha, et kui sõna võib olla oma kuju poolest nii nimisõna kui ka määrsõna, jätab mudel ühestamise tegemata.	141800
	fraas: **korraga** abielluvad õde rektsiooni võimalikud käänded: [<null>, **kom**] rektsioonil põhinev kääne: **kom** oletuslik kääne: puudub tabelis	
III	**Fraasi esimene sõna.** Analüüsitav sõna on fraasis esimesel kohal. Eesti keeles on esimesel kohal valdavalt subjekt ja subjekt omakorda valdavalt nimetavas käändes.	9067
	Pätsi vastu Laidoneri poolt agiteerinud rektsiooni võimalikud käänded: [part, <null>, gen, adit] rektsioonil põhinev kääne: part oletuslik kääne: **gen**	
IV	**Pärisnimi.** Analüüsitav sõna on fraasis suurtähestatud (ilmselt pärisnimi). Selle alajaotuse abil saab kindlaks teha kas pärisnimede ja eriti võõrapäraste pärisnimede puhul on tegemist korrektselt määratud algvormiga ja nime korrektse käändega (nt Dean-i või Dea-ni)	44992

	Ma siis aasisin **Deani** rektsiooni võimalikud käänded: [<null>, nom, term, **part**, gen, adit] rektsioonil põhinev kääne: **part** oletuslik kääne: term	
V	**Nimisõnasubjekt.** Uuritavas fraasis esineb sõltuvussuhtes nimisõnasubjekt ja võimalike käänete hulgas partitiiv. Sellise alajaotuse kasutamine on mõistlik kuna nimisõnasubjekt on eesti keeles enamasti nimetavas käändes aga reeglipõhine ühestaja on siin pakkunud osastavat käänet.	84420
	aasis **õde** rektsiooni võimalikud käänded: [**part**, nom] rektsioonil põhinev kääne: **part** oletuslik kääne: nom	
VI	**Määratlemata.** Kõik ülejäänud kirjed, mis ei mahtunud ühtegi ülalloetletud alamhulka.	357405
	aasib Leinatamm **apsu** üle rektsiooni võimalikud käänded: [part, gen, adit] rektsioonil põhinev kääne: part oletuslik kääne: **gen**	

Tabel 9. Alajaotuse valiku kirjeldused, alamhulkade suurused ja näited

Esmane alamhulga valideerimine

Algandmetest ülevaate saamiseks on mõistlik need käsitsi üle valideerida. Selleks valisime igast konflikti alamhulgast 100-kirjelise juhuvalimi ja andsime hinnangu, kas eksis reeglipõhine mudel või oletuspõhine mudel või mõlemad. Kuna valimis on ainult 100 elementi, siis on tulemused antud 5% täpsusega. Esimese alamhulga puhul saab juba olemasolevat reeglipõhist klassifitseerijat kasutada, ülejäänute puhul tuleb suures osas märgendada instrueeritavate keelemudelite abil või käsitsi (vt Tabel 10).

Alamhulk	Reeglipõhine õige	Oletuslik õige	Käsitsimärgendatud kääne on võimalike käänete hulgas
I	95%	5%	95%
II	5%	5%	95%

III	60%	10%	~100%
IV	55%	45%	90%
V	75%	25%	~100%
VI	30%	60%	~100%

Tabel 10. Käsitsi valideeritud õigsusprotsendid, alamhulkade lõikes.

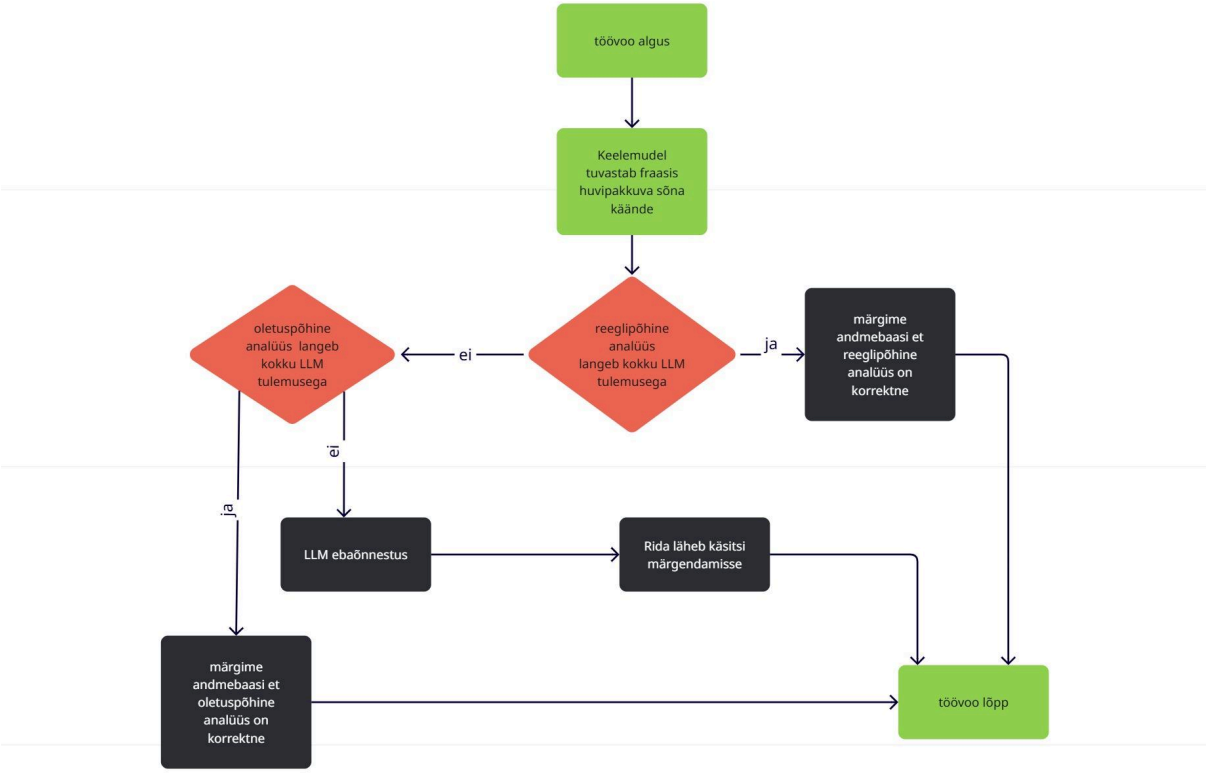
Käsitsivalideeritud valimi uurimisel tulevad ilmsiks juhud, milles oletuslik mudel eksis, aga fraasipõhine mitte (näiteid vaata Tabelit 11).

Alamhulk	Reeglipõhine õige	Oletuslik õige
I	**Sportlasi** (part/nom) ahistas seegi	**Riisalu** (nom) aitas tulla **Käärtile** (gen) järgnenud
II	ahistanud **end** (part/-) programmiga	**Hindasid** (nom) hoiab kasvamast **Jäärasid** (nom) ähvardab täna tujukus
III	**Valgust** (part/nom) aitaks heita	**Primakovi** (gen) meelest aitaks tugevdada **Eestisse** (ill) on akrediteerinud suursaadiku Helsingist Argentina
IV	sa abiellud **Egoniga** (kom/nom)	mees on meeskondadesse aidanud teiste hulgas **Kimi** (gen) aasisid muusikud **Justamendist** (el) sai
V	aitab **flamest** (el/nom)	nädalal agiteerivadki **parteid** (nom) põhiliselt Ida-Saksamaal **Ema** (nom) agiteeris hääletama
VI	angerjad ahistasid **karjabrigadiri** (part/gen)	**Raha** (gen) abil aitas ta säilitada Vaktsiin aitab **tüve** (gen) vastu

Tabel 11. Iga alamhulga kohta on esitatud 2 juhtu, mille korral tuvastati korrektselt oletuspõhiselt ja 1 juht, mis tuvastati oletuspõhiselt valesti ja reeglipõhiselt korrektselt. Reeglipõhiselt korrektselt tuvastatud kääne on allajoonitud ja kaldkriipsu järel on toodud oletuspõhiselt valesti tuvastatud kääne.

Instrueeritavate keelemudelite kasutamine nimisõna käände määramiseks

Käsitsi märgendamise osaliseks asendamiseks automaatmärgendamisega tuleb iga kirje korral järgida protsessi, mille alusel otsustatakse kas tulemuseni on võimalik jõuda instrueeritud keelemudeleid kasutades või tuleb ikkagi käsitsi märgendada (vt Joonis 3).



Joonis 3. Töövoodiagramm kirjeldab protsessi kuidas on võimalik inimmärgendamist vähendada.

Kuna vabamorf eksib nimisõna käände morfoloogilisel ühestamisel, siis on vaja alamhulkade III-VI lahendamisel teist sõltumatut märgendust. (vt Tabel 11). Instrueeritavatelt keelemudelitelt (LLM-delt), näiteks ChatGPT-4o, pole mõtet otse küsida sõnade morfoloogilist analüüsi, kuna nende treenimisel kasutatavates alusandmetest vastavaid näiteid ei ole ja seega ei ole põhjust arvata, et mudel oskaks otsesele küsimusele korrektselt vastata. Seda näitavad ka praktilised katsed. Samas LLM-id peavad teadma sõnade morfoloogiat implisiitselt, sest muidu ei oskaks nad korrektseid lauseid sünteesida. Esmane alternatiivne idee oli keelemudeli instrueerimine, nii et huvipakkuv sõna asendataks küsisõnaga, mis on eesti keeles üheselt lahutuvad. Näiteks “poiss lõi palli” puhul asendada sõna “poiss” küsisõnaga. See lähenemine andis mõningaid tulemusi, aga esiteks eksisid instrueeritavad

keelemudelid massiliselt elus/elutu kategooriaga, hakkasid kohati uusi küsisõnu välja mõtlema ja eksisid ka käändega. Järgmisena valisime asesõna “see”, mille morfoloogilised vormid lahutuvad samuti unikaalselt ja seega võimaldavad üheselt määratletavat analüüsi. Tulemus oli parem, väljamõeldud sõnade probleemi ei esinenud ja paranes ka käände tabavuse protsent. Selle metoodika hetkel teadaolevaks puuduseks on juhtumid, mille korral analüüsitavaks sõnaks on nimisõna asemel sattunud määrsõna.

Keelemudelite instrueerimisel kasutatakse teiste hulgas järgnevaid viise:

- zero-shot prompting, mille korral mudelilt küsitakse midagi ilma eelneva näiteta;
- few-shot prompting, mille korral mudel saab mõned üksikud näited (tavaliselt 1–5) enne kui ta peab ülesannet täitma ja mudel üldistab nende põhjal.

Keelemudeli instrueerimisel oli kasutuses *zero-shot prompting*. Lõpliku sobiliku prompti leidmiseks oli kasutuses eksperimenteerimine katseeksituse meetodil ja valideerimine väga väikeste valimite peal (25 kirjet), mille abil sai kiiresti sõeluda välja lähenemised, mis tegid palju vigu. ChatGPT-4o tulemus ei paranenud märkimisväärselt inglisekeelset prompti kasutades, küll aga Llama 3.1 puhul (vt Tabel 12).

Mõlemad instrueeritavad keelemudelid unustavad aegajalt nõude “tagasta ainult see sõna”, tagastades pika teksti selgitustega. Lõpliku prompti väljavalimisel sai see oluliseks kriteeriumiks, et tagastataks üks ja ainult üks sõna suure töökindlusega. Esialgne plaan iga alamhulga piires prompti optimeerida märkimisväärseid tulemusi täpsuses ei andnud, seega läks kasutusse üks universaalne prompt kõigile alamhulkadele. Lõpliku täpsuse saamiseks alamhulga lõikes valideerisin käsitsi 100 realised juhuvalimid 5% täpsusega.

Mudel	ChatGPT-4o
Sisend	*poiss lõi valge palli üle aia*
Prompt	`asenda sõna nr. 4 samas käändes ainsuses asesõnaga (see,selle,seda, jne) ja säilitades käände . tagasta ainult see sõna: poiss lõi valge palli jalaga vastu seina`
Väljund	`selle`

Tabel T. Lõplik prompt ChatGPT-4o jaoks, mis andis parimaid tulemusi.

Mudel	Llama 3.1
Sisend	*poiss lõi valge palli üle aia*
Prompt	`"replace word nr. 4 with a pronoun ('see', 'selle', 'seda', etc.) in the same case and return only that word: poiss lõi valge palli üle aia`
Väljund	`selle`

Tabel 12. Töö käigus selgus, et Llama 3.1 katses olnud valimil instrueeris ~12% täpsemini, kui teda instrueerida inglise keeles.

Töö praktiline osa on teostatud programmeerimiskeeles Python.

```
1. $mudel ChatGPT-4o
2. Loe $fraas konfliktide tabelist
3. Loe $juursõna_positsioon konfliktide tabelist
4. Küsi instrueeritavalt keelemudelilt väljund instrueerides -
   “replace word nr. $juursõna_positsioon with a pronoun ('see',
   'selle', 'seda', etc.) in the same case and return only that
   word: $phrase”
5. Kirjuta $keelemudeli_väljund konfliktide tabelisse tulemuse
   veergu
```

Pseudokood kasutusel olnud pythoni koodi peamisest osast.

Instrueeritavate keelemudelite täpsus eri alamhulkades

Selleks, et hinnata instrueeritavate keelemudelite edukust ja valida kahest alternatiivist parem, tuli iga alamhulk märgendada kahe erineva mudeliga, seejärel valideerida mõlema mudeli tulemused väikese juhuvalimi põhjal ning säilitada selle mudeli märgendused, mis osutus täpsemaks.

Riistvara ja tarkvara konfiguratsioon

Praktilise katse konfiguratsioon hõlmas nii riist- kui tarkvarakomponente. Llama 3.1 mudelit jooksutati Tartu Ülikooli Teadusarvutuskeskuse HPC-klastris. ChatGPT-4o kasutati Google Colabi pilveteenuses OpenAI API kaudu ning osaliselt ka lokaalselt sülearvutis. Mõlema keelemudeli instrueerimiseks kasutati programmeerimiskeelt Python. Andmete salvestamiseks ja päringute teostamiseks kasutati kergekaalulist relatsioonilist andmebaasi SQLite.

Esmased tulemused ja nende töötlemine.

Toetudes kirjeldatud töövoodiagrammile (vt Joonis 3) tuleb iga alamhulga piires jagada kirjed kolmeks. Juhul, kui IKM analüüs ühtib oletusel põhineva analüüsiga (rida 1, roheline) võime antud kirja eraldada potentsiaalselt lahendatud konfliktide hulka. Juhul, kui IKM analüüs ühtib rektsioonianalüüsiga (rida 2, kollane) võime antud kirje eraldada oletusel põhineva analüsaatori edasiste treeninganmdete hulka. Juhul, kui IKM pakutav kääne ei lange kokku rektsioonil põhineva analüüsiga ega oletuspõhise analüüsiga ja seda käänet ei ole ka võimalike käänete hulgas, siis on suure tõenäosusega kääne korrektselt määramata ja tuleks edasi saata käsitsimärgendamisse (vt Joonis 4).

Alamhulk III					
rida	fraas	analüüsitav sõna	rektsioonil põhinev analüüs	oletusel põhinev analüüs	IKM analüüs
1	Pätsi vastu Laidoneri poolt agiteerinud	Pätsi	part	**gen**	**gen**
2	Kivisildnikku ahistanud	Kivisildnikku	**part**	adit	**part**
3	Valgust aitavad peegeldada	Valgust	el	nom	**part**

Alamhulk I alajaotus 1 - Kuna vabamorf analüüsib ka IKM hinnangul korrektselt siis võib selle fraasi eraldada potentsiaalselt lahendatud konfliktide hulka.					
rida	fraas	analüüsitav sõna	rektsioonil põhinev analüüs	oletusel põhinev analüüs	IKM analüüs
1	Pätsi vastu Laidoneri poolt agiteerinud	Pätsi	part	**gen**	**gen**

Alamhulk I alajaotus 2 - Kuna rektsioonil põhinev analüüs ühtib IKM-i analüüsiga. siis saab seda fraasi kasutada oletusel põhineva analüsaatori täiendaval treenimisel.					
rida	fraas	analüüsitav sõna	rektsioonil põhinev analüüs	oletusel põhinev analüüs	IKM analüüs
2	Kivisildnikku ahistanud	Kivisildnikku	**part**	adit	**part**

Alamhulk I alajaotus 3 - Kuna IKM-i analüüs ei lange kummagi analüüsiga kokku, siis järelikult see alamhulk peab minema käsitsi märgendamisse.						
rida	fraas	analüüsitav sõna	rektsioonil põhinev analüüs	oletusel põhinev analüüs	võimalikud käänded	IKM analüüs
1	Valgust aitavad peegeldada	Valgust	el	nom	**[el]**	**part**

Joonis 4. Näide IKM hinnangu põhjal kolme alamhulka jaotatud kirjetest.

Analüüsitav sõna on fraasis suure tähega. Selles alamhulgas on pärisnimed, mille morfoloogilise muutumise ennustamine on keskmisest raskem ülesanne, eriti kui tegemist on võõrapäraste nimedega. Manuaalse validatsiooni tulemuste põhjal oli reeglipõhise ühestaja täpsus 53% ja oletuspõhise ühestaja täpsus 35%. Ülejäänud 12%-l juhtudel eksisid mõlemad mudelid.

Eksprimendi disain

Katse eesmärk on märgendada kahe erineva instruueritava keelemudeli abil sõnade käändeid, ja käsitsi valideerida märgenduste õigsus käsitsi 100-realisel juhuvalimil. Seejärel hinnata, kas instrueeritava keelemudeli määratud kääne vastab tegelikkusele, ning arvutada tabavusprotsent.

Sama Pythoni kood tuli käivitada nii ChatGPT-4o kui ka Llama 3.1 peal märgendamiseks. Käsitsi valideerimise alusel 100 realise juhuvalimi peal valisin igas alamhulgas parima mudeli. Enamikul juhtumistest oli ChatGPT-4o parem.

Alamhulk konfliktide andmebaasis	ridu	ChatGPT-4o tulemus	Llama 3.1 tulemus
1. Vaid üks võimalik kääne	17322	**85%**	80%
2. Võimalike käänete hulgas null ja veel midagi.	141800	**25%**	20%
3. Esimene sõna uuritavas fraasis	9067	**78%**	64%
4. Uuritav sõna on suurtähestatud ehk pärisnimi.	44992	41%	**43%**
5. sõltuvussuhe NSubj, fraasipõhise analüüsi kääne partitiiv, oletusel põhineva analüüsi tulemus nimetav (konflikt)	84420	**50%**	42%
6. kõik ülejäänud kirjed	357407	**62%**	35%

Tabel 13. Alamhulgaks määramise kriteeriumid, suurus ridade arvuna, alamhulkade LLM-i abil märgendamise täpsusprotsent, üldine täpsusprotsent (55%). Sõnad märgendati kahe erineva IKM-iga - ChatGPT-4o ja LLaMA3.1 ja valiti peale käsitsi valideerimist neist parim.

Puudujäägid ja edasised suunad

Konfliktseid kirjeid instrueeritavate keelemudelite abil kategoriseerides on võimalik ühe lisasammuna teostada puuduolevate rektsioonimallide tuvastamine juhtudel, kus instrueeritava keelemudeli vastus on analüüsitava sõna võimalike käänete hulgas ja see langeb kokku Vabamorfi analüüsiga. Näiteks "lööma" verbil on andmebaasis rektsioonimall “lööma kuhu”. Sellistel juhtudel on võimalik automaattuvastada lisaks rektsioonid "lööma mida" või "lööma millega".

Töös kasutuses olnud metoodikale toetudes on võimalik koostada märgendamisülesanded inimestele juhtudel, kus instrueeritava keelemudeli tulemus ei ole võimalike käänete hulgas ja ei lange ka kokku reeglipõhise ega oletusel põhineva analüüsiga.

Samuti jääb probleemiks automaatse süntaksianalüüsi vigade tuvastamine. Seda on võimalik teha, vaadates ühe verbi kõiki võimalikke rektsioonimalle, kuid see ei pruugi anda soovitud tulemust, kuna mõned rektsioonid on väga haruldased. Lisaks puudub hetkel ka rektsioonimallide andmebaas, mis sisaldaks isegi väikse hulga verbide kõiki rektsioonimalle.

Tulemused

Kaasaegsete morfoloogilise ja süntaktilise analüüsi meetodite arendamine nõuab kvaliteetseid ja märgendatud andmeid, mille kogumine on kallis ja aeganõudev. Täpsuse parandamine üle 90% muutub üha raskemaks ja kulukamaks. Käesolev töö keskendub rektsioonide rollile analüüsivigade tuvastamisel, uurides morfoloogia ja süntaksi kooskõla. Eesmärk on vähendada käsitsi andmetöötlust, arendades automaatseid meetodeid vigade leidmiseks ja parandamiseks, kasutades instrueeritavaid keelemudeleid.

Töös kasutati alusena andmebaasi, kuhu on koondatud konfliktid morfoloogilise ja süntaktilise analüüsi vahel, mis ilmnevad tänaste automaatsete tööriistade kasutamisel. Need konfliktid viitavad juhtumitele, kus sõnade morfoloogilised tunnused ei vasta süntaktilisele struktuurile või vastupidi. Et analüüsi süstematiseerida, jaotati konfliktid väiksemateks alamhulkadeks vastavalt nende keelelisele ja struktuurilisele iseloomule.

Edasi kasutati kahte instrueeritavat keelemudelit, et ennustada, milline võiks olla keeleliselt korrektne analüüs igas konfliktis. Alles jäeti ainult need juhud, kus mõlemad mudelid pakkusid sisuliselt ühesuguse lahenduse. Seejärel viidi läbi käsitsi kontroll valimi põhjal, et hinnata sellise lähenemise täpsust ja usaldusväärsust. Kontrolli tulemused võimaldasid arvutada täpsusprotsendi ja teha järeldusi meetodi kasutatavuse kohta automaatses andmekvaliteedi tõstmises.

Teatudel juhtudel andis metoodika küllaltki häid tulemusi. Näiteks alamhulkade I (vaid üks võimalik kääne - 87% õigesti üle märgendatud) ja III (esimene sõna uuritavas fraasis - 78% õigesti üle märgendatud). Neil juhtudel oli fraasidel enamasti väga lihtne ja üheselt mõistetav ehitus. Alamhulk I analüüsiga sai rektsioonipõhine analüsaatoor paremini hakkama kui instrueeritav keelemudel, seega on see osa juba algandmete tabelis lahendatud.

Alamhulk II (Võimalike käänete hulgas <null> ja veel 1 kääne) sisaldas suures osas süntaksikihist vigadena sisse tulnud määrsõnu. Määrsõnad on eesti keeles muutumatud. Rektsioonil põhinev analüsaator püüab ka määrsõnu analüüsida kui nimisõnu, näiteks “tavaliselt”, sg all. Võimalik, et instrueeritavalt keelemudelilt saab küsida kas sõna on määrsõna. Esmased katsetused kinnitasid, et see nii on ja seda on võimalik edasi arendada.

Ülejäänud alamhulkades jäi täpsus vahemikku 45 - 64%. Näiteks “aasis õde (venda)” - nominatiiv, aga “(vend) aasis õde” - partitiiv. Täpselt üks fraasi kuju võib vastata mitmele

alternatiivsele analüüsile ja mõlemad on korrektsed. Sellisel juhul võib olla tegemist juhuga kus on liiga vähe konteksti kaasa võetud algsest lausest ja üks võimalik lahendus anda sellistel juhtudel algne täispikk lause instrueeritud keelemudelile sisendiks. Ilmselt on võimalik õigsusprotsenti keelemudelite osava instrueerimise abil veel parandada, sest kohati tulevad sisse pealtnäha lihtsad vead. Näiteks ChatGPT-4o on analüüsinud "selgitustööd aitab teha", selgitustöö, sg n. "Selgitustööd" saaks olla siin sg p või pl n, aga mitte sg n. Sellistele juhtumitele ei õnnestunud autoril lahendust leida. hinnanguliselt oli neid suurusjärgus ~10%.

Alamhulka VI (täpsus ~62%) täpsemalt uurides on näha, et ka sellesse alamhulka on sattunud päris palju süntaksikihis valesti määratud sõnu - määrsõnu, tegusõnu jne (~27000). Näiteks "ütles korduvalt kuule" - verb *ütlema,* abalüüsitav sõna *kuulma*.

Samuti on antud töö peale võimalik edasi ehitada näiteks LabelStudio märgendamisülesanne.

Viited

[1] Raili Pool. Eesti keele verbirektsioone. Tartu: Tartu Ülikooli Kirjastus, lk 5, 1996.
https://dspace.ut.ee/server/api/core/bitstreams/749e7929-fe79-4d60-8482-08c1f730243a/content

[2] Heiki-Jaan Kaalep,Tarmo Vain. Vabamorf. 2001
https://github.com/Filosoft/vabamorf

[3] Heiki-Jaan Kaalep jt . ESTNLTK - NLP Toolkit for Estonian . 2016. Tartu Ülikool. 2016
http://www.lrec-conf.org/proceedings/lrec2016/pdf/332_Paper.pdf

[4] Heiki-Jaan Kaalep, Sven Laur, Alexander Tkachenko, Timo Petmanson. EstNLTK Tables of morphological categories. Tartu. Tartu Ülikool. 2016.
https://github.com/estnltk/estnltk/blob/main/tutorials/nlp_pipeline/B_morphology/00_tables_

[5] Heiki-Jaan Kaalep, Tarmo Vaino. Complete Morphological Analysis in the Linguist's Toolbox. Tartu. Tartu Ülikool. 2001
https://www.cl.ut.ee/yllitised/smugri_toolbox_2001.pdf

[6] Kertu Saul. EstNTLK morfoloogilise analüsaatori ja ühestaja kvaliteedi hindamine. Tartu Tartu Ülikooli Kirjastus. 2022
https://dspace.ut.ee/server/api/core/bitstreams/89e22ec0-b674-45b9-b117-edc5fec843b1/content

[7] Daniel Jurafsky jt. Speech and Language Processing. Stanford University, lk 395, 2024
https://web.stanford.edu/~jurafsky/slp3/ed3book_Jan25.pdf

[8] Laura Katrin Leman. Tehisnärvivõrgul põhinevate lemmatiseerijate võrdlev analüüs eesti keeles. Tartu. Tartu Ülikool, lk 19, 2019
https://dspace.ut.ee/server/api/core/bitstreams/3b0bd18a-6b74-44f3-bc74-5edd9ab4d4c3/content

[9] Sven Laur, Siim Orasmaa, Sandra Eiche, Dage Särg. Automatic dependency parsing of Estonian: what linguistic features to include?, lk 10, 2024.

https://scholar.google.com/citations?view_op=view_citation&hl=en&user=663HUJIAAAAJ&sortby=pubdate&citation_for_view=663HUJIAAAAJ:nb7KW1ujOQ8C

[10] Sirje Mäearu. Valik rektsioone. Tartu: Tartu Ülikooli Kirjastus, lk 3, 2011. https://www.digar.ee/arhiiv/en/download/107743

[11] Langemets, Margit; Hein, Indrek; Jürviste, Madis; Kallas, Jelena; Kiisla, Olga; Koppel, Kristina; Kuusk, Külli; Leemets, Tiina; Mäearu, Sirje; Paet, Tiina; Päll, Peeter; Raadik, Maire; Risberg, Lydia; Rehemaa, Tuuli; Tiits, Mai; Tsepelina, Katrin; Tuulik, Maria; Uibo, Udo; Valdre, Tiia; Viks, Ülle ... Volkova, Dona (2024). EKI ühendsõnastik 2024. Eesti Keele Instituut.
Väljavõte EKI Ühendsõnastikust 21.08.24 seisuga: eesti ilmik rektsiooniga.
https://github.com/estnltk/syntax_experiments/blob/verb_templates/workflows/004_analysis_of_known_verb_rection_patterns/source_data/12_eesti_ilmik_rektsiooniga_linkideta.csv

[12] Raili Pool. Mis on sihitis ? Tartu. Eesti ja üldkeeleteaduse instituut. 2013
https://sisu.ut.ee/sihitis/1-mis-sihitis/

[13] Eesti Keele Instituut. Eesti keele koondkorpus. 2018
https://cl.ut.ee/korpused/segakorpus/index.php?lang=et
of_morphological_categories.ipynb

[14] Katrin Tsepelina, Sven Laur. Koondkorpuse verbifraaside andmebaas.
https://github.com/estnltk/syntax_experiments/tree/verb_templates

[15] Katrin Tsepelina, Sven Laur. Võimalike morfoloogia ja süntaksi konfliktide andmestik.
https://github.com/estnltk/syntax_experiments/tree/verb_templates/workflows/004_analysis_of_known_verb_rection_patterns/004_extracting_syntax_morphology_conflicts/example_data

[16] High Performance Computers and Services / Rocket.
https://hpc.ut.ee/services/HPC-services/Rocket

Lihtlitsents

lõputöö reprodutseerimiseks ja lõputöö üldsusele kättesaadavaks tegemiseks

Mina, Hendrik Aruoja (sunnikuupaev: 04.06.1983)
1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) enda loodud teose „Keeletehnoloogiline abivahend, mille abil saab tuvastada verbimalle ja leida nimisõnade semantilisi klasse.“, mille juhendaja on Sven Laur PhD.
1.1. reprodutseerimiseks säilitamise ja üldsusele kättesaadavaks tegemise eesmärgil, sealhulgas digitaalarhiivi DSpace-is lisamise eesmärgil kuni autoriõiguse kehtivuse tähtaja lõppemiseni;
1.2. üldsusele kättesaadavaks tegemiseks Tartu Ülikooli veebikeskkonna kaudu, sealhulgas digitaalarhiivi DSpace´i kaudu alates 01.01.2026 kuni autoriõiguse kehtivuse tahtaja lõppemiseni.
2. olen teadlik, et nimetatud õigused jäävad alles ka autorile.
3. kinnitan, et lihtlitsentsi andmisega ei rikuta teiste isikute intellektuaalomandi ega isikuandmete kaitse seadusest tulenevaid õigusi.
Tartus, 15.05.2025

Lisad

Käesoleva lõputöö koodirepositioorium

https://github.com/gitpandasmail/languagetechnologicaltool