Tartu Ülikool
Maailma keelte ja kultuuride instituut

Marii Ojastu

WINOGRANDE ANDMESTIKU TÕLKIMINE SUURTE KEELEMUDELITE ARGIMÕISTUSLIKU JÄRELDAMISOSKUSE HINDAMISEKS EESTI KEELES

Magistritöö

Juhendajad:
Kairit Sirts, PhD
Marika Borovikova, kirjaliku tõlke nooremlektor

Tartu
2025

SISUKORD

SISSEJUHATUS

Mitmed vabavaralised keelemudelid suudavad juba eesti keelt mõista, kuid Eesti keeletehnoloogiate arendamine on riiklikul tasandil jätkuvalt oluline. Seda kinnitab ka Haridus ja Teadusministeeriumi programm „Eesti keeletehnoloogia 2018–2027“, mis toetab keeletehnoloogiaalast teadus- ja arendustööd. Programmi eesmärk on luua uusi eestikeelseid keeletehnoloogia rakendusi, tõsta olemasolevate rakenduste kvaliteeti ning võtta neid kasutusele võimalikult laialdaselt eri valdkondades – nii era-, avalikus kui ka kolmandas sektoris. Programmi raames arendatakse keeletehnoloogiate põhikomponente, et need saavutaksid rahvusvahelises võrdluses hea taseme ning programmi tulemusena nähakse eesti keeletehnoloogiate kasutamist laia sihtgrupi poolt (Haridus- ja Teadusministeeriumi kodulehekülg 2024). Programmi teadustegevuses osaleb ka Tartu Ülikooli arvutiteaduse instituut, mille juhitud projekti raames õpetatakse suurtele keelemudelitele eesti keelt ja kultuuri. Keelemudelite treenimise eesmärk on eesti keele ja kultuuri säilitamine ja kaitsmine tehisaru kiire arengu ajastul ning nende keelemudelite kasutamine erinevates rakendustes, mida eestlased edaspidi kasutada saavad (*ERR News*, 04.09.2024 kl 16.04).

Keelemudelite oskusi ja teadmisi mingis kindlas keeles on võimalik kvantitatiivselt hinnata. Selleks otstarbeks on välja töötatud mitmeid mahukaid testandmestikke, mis enamasti koosnevad erinevat tüüpi tekstülesannetest. Iga testandmestik on koostatud mingi konkreetse oskuse või konkreetsete teadmiste hindamiseks. Selliseid testandmestikke avaldatakse reeglina suure kõnelejaskonnaga keeltes, näiteks inglise keeles. Väiksema kõnelejaskonnaga keelte tarbeks on neid andmestikke masintõlgitud, kuid selline lähenemine pole alati tõhus, kuna masintõlke kvaliteet mõjutab keelemudelite testisooritust (Thellmann jt 2024). Selle vältimiseks on soovitatav tõlke toimetamine ekspertide poolt ja tõlgitud andmestike kohandamine lähtekeele eripäradega (Plaza jt 2024). Olemas on ka mitmekeelseid testandmestikke, mis hõlmavad eesti keelt (nt Bandarkar jt 2023; Ponti jt 2020), kuid selliste andmestike osakaal on väike.

Üks keelemudelite oskuste kvantitatiivseks hindamiseks koostatud andmestik on ingliskeelne andmestik nimega WinoGrande. WinoGrande andmestik koosneb anafooride ehk tekstisiseste tagasiviidete lahendamise ülesannetest ning selle andmestikuga saab hinnata keelemudelite argimõistuslikku järeldamisoskust (inglise keeles *commonsense reasoning*). Anafooride lahendamine on masintõlkemootoritele ning ka keelemudelitele jätkuvalt keeruline ülesanne (Naveen ja Trojovský 2024). Kuna masintõlkemootorid ei tule anafooriliste viitesuhete

tõlkimisega veel hästi toime, võib eeldada, et WinoGrande andmestikku ei saa eesti keelde masintõlgituna eesmärgipäraselt kasutada.

Käesoleva magistritöö eesmärk on ingliskeelne WinoGrande testandmestik eesti keelde tõlkida, lokaliseerida ja eesti keelele kohandada. Magistritöös tõlgitakse 1767 tekstülesannet, mille kogumaht on 37 802 sõna. Tegemist on omapärase tõlkega, kuna tõlke lugeja ei ole inimene, vaid masin. Magistritöös tuvastatakse ja dokumenteeritakse andmestiku eesti keelde tõlkimises esinevad väljakutsed ja kirjeldatakse meetodeid, mis toetaksid taoliste andmestike eesti keelde tõlkimist ka tulevikus. Magistritöö raames valminud andmestikku kasutatakse Tartu Ülikooli arvutiteaduse instituudis eestikeelsete keelemudelite järeldamisoskuse hindamiseks ning saadud tulemusi kajastatakse ka selles magistritöös. Kuna taolisi andmestikke on teistesse keeltesse samal eesmärgil ka masintõlgitud, siis analüüsitakse magistritöös lisaks seda, kas eestikeelset masintõlget oleks võimalik keelemudelite hindamiseks kasutada. Keelemudelid, mille järeldamisoskust eesti keeles hinnatakse, on OpenAI GPT-4o, EuroLLM 9B (Martins jt 2024), Llammas (Kuulmets jt 2024), LLama 3.3 70B (Grattafiori jt 2024), LLama 3.1 8B (Grattafiori jt 2024) ja LLama 3.1 405B Instruct (Grattafiori jt 2024). Tõlgitud andmestikku saab kasutada eesti keeletehnoloogia arendamises ka edaspidi.

Magistritöö teoreetilises osas antakse ülevaade WinoGrande testandmestikust ja selle koostamise aluseks olnud põhimõtetest. Lisaks kirjeldatakse tõlkeprobleeme, mida teised autorid on WinoGrande andmestikku erinevatesse keeltesse tõlkides täheldanud. Töö empiirilise osa esimeses pooles kirjeldatakse andmestiku eesti keelde tõlkimisel tuvastatud tõlkeprobleeme ning analüüsitakse, kuidas masintõlge nende tõlkeprobleemidega toime tuleb. Empiirilise osa teises pooles esitatakse keelemudelite testisoorituse tulemused ingliskeelse andmestikuga, eestikeelse masintõlgitud andmestikuga ning käesoleva magistritöö raames valminud eestikeelse tõlgitud ja lokaliseeritud andmestikuga.

1. TEOREETILIS-METODOLOOGILINE RAAMISTUS

1.1 Suured keelemudelid ja sõnavektorid

Eesti Keele Instituudi teatmik defineerib suuri keelemudeleid (SKM) järgnevalt: „SKMid on arvutiprogrammid, mis on treenitud ennustama eelnevate sõnade põhjal järgmist sõna. SKMid õpivad keelt andmekogude põhjal, märgates mustreid sõnade, lausete ja isegi terviktekstide vahel.“ (Eesti Keele Instituudi teatmik 2025). Seesugune mustrite märkamine ja ennustamine ei toimu tänapäeva suurtes keelemudelites mitte sõnade enda, vaid sõnu ja nende semantilisi seoseid esindavate reaalarvude jadade ehk vektorite kaudu. Selliseid vektoreid nimetatakse sõnavektoriteks (inglise keeles *embeddings*). Sõnavektorites väljendub sõnadevaheline lingvistiline seos, mis on tuletatud suuri tekstikorpusi matemaatiliselt analüüsides. Tähenduslikult sarnastel sõnadel on sarnased sõnavektorid ja sõnavektorite kaudu on võimalik sõnadevahelist semantilist sarnasust hinnata (Saleh ja Paquelet 2024). Tehisnärvivõrkudes toimuvate keerukate matemaatiliste korrutuste tulemusena luuakse sõnavektoritest sõnajadade konteksti esindavad vektorid, milles väljendub konteksti semantiline tähendus. Kui tekstid kasutavad erinevaid sõnu, kuid nende tähendus on sarnane, siis on tekstidel sarnased reaalarvulised vektorid (Sirts 2024).

1.2 Keelemudelite oskuste ja teadmiste hindamine

Keelemudelite oskuste ja teadmiste kvantitatiivseks hindamiseks on välja töötatud mitmeid mahukaid andmestikke, mis enamasti koosnevad erinevat tüüpi tekstülesannetest, millele keelemudel valib vastusevariantide hulgast õige vastuse. Käesoleva magistritöö kontekstis nimetatakse selliseid andmestikke testandmestikeks (inglise keeles *benchmark dataset*). Üldtuntud testandmestikest on hiljuti avaldanud ülevaate Ivanov ja Penchev (2024), kes on oma töös loetlenud 40 seesugust andmestikku. Selliseid andmestikke on rohkem, kuid kuna puudub ühtne valdkonnaülene süsteem, mille kaudu selliseid andmestikke avaldatakse (Longjohn jt 2024), on nende andmestike täpset arvu raske hinnata. Üks testandmestik võib koosneda mitmekümnest tuhandest samalaadsest ülesandest, mis on välja töötatud keelemudeli spetsiifilise oskuse, omaduse või üldiste teadmiste kvantitatiivseks hindamiseks.

Seesuguste testandmestike väljatöötamine on töömahukas protsess ning need andmestikud võivad keelemudelite kiire arengu tõttu juba mõne aastaga kasutuks muutuda. Üldiselt avaldatakse neid andmestikke suure kõnelejaskonnaga keelte tarbeks ning väiksema kõnelejaskonnaga keelte tarbeks on neid andmestikke masintõlgitud (Plaza jt 2024). Kui andmestikke masintõlgitakse, on testi tulemus moonutatud, sest tulemuses kajastub nii keelemudeli sooritus kui ka tõlke kvaliteedi mõju sellele sooritusele. On näidatud, et masintõlke järeltoimetamine ja lokaliseerimine parandab oluliselt testi tulemuse kvaliteeti (*ibid*.).

Ideaalolukorras oleks testandmestik konkreetsele keelele kohandatud või koostatud selles keeles, milles keelemudelit testitakse (Plaza jt 2024). Testandmestike tõlkimisel on oluline jälgida, et andmestiku eesmärk oleks tõlkes säilitatud. Andmestike tõlkimine spetsialistide poolt muudab tulemused selgemini tõlgendatavaks ja ka võrreldavaks, kuna tõlke kvaliteedi mõju tulemustele on sel juhul eeldatavasti väiksem.

1.3 WinoGrande andmestik

Üks keelemudelite oskuste kvantitatiivseks hindamiseks koostatud andmestik on ingliskeelne andmestik nimega WinoGrande. WinoGrande andmestik koosneb anafooride lahendamise ülesannetest ning selle andmestikuga saab hinnata keelemudelite argimõistuslikku järeldamisoskust (Sakaguchi jt 2019). Anafoorid on tekstisisesed tagasiviited (Pajusalu 2017: 567) ja nende lahendamine on masintõlkemootoritele ning ka keelemudelitele jätkuvalt väljakutsuv (Naveen ja Trojovský 2024). Kuna masintõlkemootorid ei tule anafooriliste viitesuhete tõlkimisega veel hästi toime, võib eeldada, et WinoGrande andmestikku ei saa masintõlgitud kujul keelemudelite hindamiseks eesmärgipäraselt kasutada.

WinoGrande andmestik avaldati 2019. aastal ning tegemist on Levesque'i, Davise ja Morgensterni poolt 2012. aastal avaldatud Winograd Schema Challenge'i andmestiku edasiarendusega. Winograd Schema Challenge'i testandmestik oli selle autorite sõnul alternatiiv Turingi testile. Andmestik koosnes spetsialistide koostatud 273-st tekstülesandest, mille lahendamine on inimeste jaoks lihtne, aga statistiliste keelemudelite jaoks keeruline (Sakaguchi jt 2019). Keeletehnoloogia on viimase kümnendi jooksul teinud läbi märkimisväärse arengu ning statistilistelt keelemudelitelt on üle mindud närvivõrkudel põhinevatele keelemudelitele (Jing ja Xu 2019). Närvivõrkudel põhinevad keelemudelid suudavad

Winograd Schema Challenge'i andmestiku ülesanded juba 90%-lise täpsusega lahendada, kuid see ei pruugi demonstreerida nende argimõistuslikku järeldamisoskust, vaid nende ülesannete ülesehituses võib sõnavektorite tasandil esineda teatud kallutatus, mida mudel õige vastuse tuletamiseks ära kasutab. WinoGrande testandmestiku ülesanded sarnanevad ülesehituselt Winograd Schema Challenge'i andmestiku ülesannetele, kuid WinoGrande andmestikku kogutud ülesanded on kontrollitud spetsiifilise algoritmiga, mis kallutatuse tuvastab. Ülesanded, milles algoritm kallutatuse tuvastas, jäeti andmestikust välja ning sellise metoodika kasutamine muudab WinoGrande andmestiku keelemudelite jaoks keerulisemaks (Sakaguchi jt 2019).

1.4 Winogradi ülesanded

WinoGrande testandmestik koosneb Winogradi ülesannetest, mis on oma nime saanud arvutiteadlase Terry Winogradi järgi, kes 1972. aastal esimesena taolise ülesande välja pakkus (Levesque jt 2012). Winogradi ülesanne on lühike lüngaga tekstülesanne. Keelemudeli ülesanne on täita lünk sobiva nime, nimisõna või nimisõnafraasiga, mida on lauses varem mainitud. Nimi, nimisõna või nimisõnafraas esitatakse keelemudelile vastusevariantidena ning tegemist on alati anafooriga ehk tagasiviitega. Igal ülesandel on kaks vastusevarianti. 2012. aastal koostatud Winograd Schema Challenge'i andmestikus oli ülesanne esitatud alati lausepaarina. WinoGrande andmestikus on suurem osa ülesandeid esitatud üksikute lausetena (tabel 1), kuid esineb ka lausepaare (tabel 2).

Tabel 1

Winogradi ülesanne

Ülesanne	Vastusevariant 1	Vastusevariant 2
Erinevalt Jakobist ei leidnud Aare oma postkastist kunagi võõraid kirju. _ oli väga levinud perekonnanimi.	Jakobil	Aarel

Tabel 2

Winogradi ülesanne lausepaarina

Ülesanne	Vastusevariant 1	Vastusevariant 2
Auto peatus enne veokit, sest _ roolis olnud juht vajutas pidurit hiljem.	auto	veoki
Auto peatus enne veokit, sest _ roolis olnud juht vajutas pidurit varem.	auto	veoki

Winogradi ülesanded on inimeste jaoks lihtsad, kuid keelemudelite jaoks keerulised, sest ülesannete lahendamine nõuab teatud üldteadmisi maailma toimimise kohta. Näiteks peaks keelemudel ülaltoodud näidetes (tabel 1 ja 2) teadma, et levinud perekonnanimega inimeste postkasti võivad saabuda võõrad kirjad ja kiiremini peatub see sõiduk, mille pidurit varem vajutatakse. Sellist arutlusvõimet nimetatakse argimõistuslikuks järeldamisoskuseks, kuid WinoGrande andmestiku autorid seavad kahtluse alla selle, et keelemudelid sellise oskuse on omandanud (Sakaguchi jt 2019).

1.5 WinoGrande andmestiku koostamise aluseks olnud põhimõtted

WinoGrande andmestiku Winogradi ülesanded on kogutud ühisloome (inglise keeles *crowdsourcing*) teel. Ülesannete koostajatel paluti järgida Winograd Schema Challenge'i andmestiku ülesannete koostamise aluseks olnud põhimõtteid (Sakaguchi jt 2019), mille Levesque jt on 2012. aastal kokku võtnud järgnevalt.

1) Ülesandes on kaks osapoolt. Osapooled võivad olla kaks samast soost isikut, kaks eset või kaks inimeste või esemete rühma.
2) Lauses kasutatakse sõna või fraasi, mis viitab ühele osapoolele, kuid pole välistatud ka teise osapoole jaoks. Esimene vastusevariant on lauses esimesena mainitud osapool, teine vastusevariant on lauses teisena mainitud osapool. Keelemudeli ülesanne on tuvastada osapool, kellele või millele viidatakse.
3) Lausepaarid sisaldavad ühte erilist sõna või fraasi. Kui see sõna või fraas ära vahetada, säilib lause loogilisus, kuid õige vastus muutub (vt tabel 2).

4) Ülesande lahendust ei tohiks saada tuletada ainuüksi selle põhjal, et osapoolele viitav sõna asub osapoolest semantilises tähendusvõrgustikus kaugel (vt tabel 3).
5) Ülesande vastus ei tohiks olla Google'i otsingumootori statistika põhjal lihtsasti tuletatav, kuna see tähendaks, et piisavalt suurt korpust statistiliselt analüüsides oleks vastus ilmne.

Tabel 3

Keelemudeli jaoks lihtne Winogradi ülesanne

Ülesanne	Vastusevariant 1	Vastusevariant 2
Ralliauto kihutas koolibussist mööda, sest _ sõitis väga aeglaselt.	ralliauto	koolibuss

Tabelis 3 toodud näide on keelemudeli jaoks liiga lihtne ülesanne, sest sõnad *aeglane* ja *ralliauto* asuvad semantilises tähendusvõrgustikus teineteisest oluliselt suurema tõenäosusega kaugemal kui *aeglane* ja *koolibuss* (Levesque jt 2012).

WinoGrande andmestiku laused koguti esialgu lausepaaridena, mille laused on 15–30 sõna pikad ja mille sõnadevaheline kattuvus on vähemalt 70%. Ühisloome kaudu koguti algselt 43 972 ülesannet. Kogutud ülesanded kontrolliti ning ülesanne loeti tingimustele vastavaks, kui:

1) kaks kolmest kontrollijast pidasid sama vastust õigeks;
2) kontrollijad nõustusid sellega, et üks vastustest on tõenäolisem kui teine;
3) õiget vastust ei saanud tuletada ainuüksi selle põhjal, et osapoolele viitav sõna asub osapoolest semantilises tähendusvõrgustikus kaugel (vt tabel 3).

Ülesanded, mis nendele tingimustele ei vastanud, jäeti andmestikust välja. Seejärel analüüsiti andmestikku algoritmiga, mis tuvastab sõnavektorite tasandil mustrid, mis muudavad ülesande keelemudelile lihtsaks. Kuna algoritm filtreeris lausepaaridest välja mõnel juhul vaid ühe lause, on lõplikus andmestikus kahte tüüpi ülesandeid (vt tabel 1 ja tabel 2). Kontrollimise ja filtreerimise tulemusena jäi alles 12 292 ülesannet, mis jagunevad treeningandmestikuks (9248 ülesannet), valideerimisandmestikuks (1267 ülesannet) ja testandmestikuks (1767 ülesannet) (Sakaguchi jt 2019). Selle magistritöö kontekstis on keelemudeli testimiseks vaja vaid tõlgitud testandmestikku, treeningandmestiku ja valideerimisandmestiku tõlkimine pole vajalik. Treeningandmestik ja valideerimisandmestik olid vajalikud varasemate keelemudelite tarbeks, kuid kaasaegsed keelemudelid neid andmestikke ei vaja.

1.6 Winogradi ülesannete tõlkimine

WinoGrande andmestik erineb keelekasutuse poolest Winograd Schema Challenge'i andmestikust, kuna Winograd Schema Challenge'i andmestiku ülesanded koostasid spetsialistid, kuid WinoGrande andmestiku ülesanded on kogutud ühisloome teel. Seepärast on WinoGrande andmestiku koostamisel kasutatud inglise keel mitmekülgsem ja igapäevasem (Sakaguchi jt 2019).

Winogradi ülesannete tõlkimise eripärasid on varasemalt kirjeldanud mitu autorit. Näiteks on avaldatud töid Winogradi ülesannete tõlkimisest portugali (Melo jt 2019), hiina (Bernard ja Han 2020), ungari (Vadász ja Ligeti-Nagy 2022) ja prantsuse (Amsili ja Seminck 2017) keelde. Tõlkimisel esinevad väljakutsed on seotud sihtkeele omapäradega (Davis 2016). Näiteks grammatilise sooga keelte puhul on Winogradi ülesannete tõlkimine raskendatud, sest keelemudel saab viitesuhte tuvastada artiklite ja sõnade soo järgi ning sel juhul ei ole tegemist kontekstipõhise viitesuhte tuvastamisega (Vádasz ja Ligeti-Nagy 2022). Lisaks sellele ei pruugi tõlkes säilida ülesande ühemõttelisus, kuna otsene tõlkevaste võib sihtkeeles omada mitut tähendust. Kui tõlge sisaldas probleemseid sõnu, on selliseid sõnu andmestiku tõlkimisel asendatud uute sõnadega, mis ei ole probleemsete sõnade tõlkevasted, kuid võimaldavad säilitada ülesande eesmärgi (Amsili ja Seminck 2017). Eespool loetletud autorid on oma töödes keskendunud Winograd Schema Challenge'i andmestikku kuuluvate Winogradi ülesannete tõlkimisele. WinoGrande andmestiku ülesannete tõlkimist on käsitlenud 2021. aastal Emelin ja Sennrich, kes on ülesandeid saksa, prantsuse ja vene keelde masintõlkinud. Kuna paljud ülesanded kaotavad nendesse keeltesse masintõlkimisel oma algse eesmärgi, jätsid autorid keerukamad konstruktsioonid andmestikust välja. Näiteks jäeti välja ülesanded, mille vastusevariandid olid erinevast grammatilisest soost, ning ülesanded, mille vastusevariant oli tõlkes liitsõna osa. Autorite töö tulemusena valmis kohandatud andmestik, mis kannab nime Wino-X. See sisaldab ühtede ja samade Winogradi ülesannete saksa-, vene- ja prantsuskeelseid tõlkeid ning sellega saab testida mitmekeelseid suuri keelemudeleid *(*Emelin ja Sennrich 2021).

2. WINOGRANDE ANDMESTIKU TÕLKIMINE

2.1 Andmestiku vorming

WinoGrande andmestiku saab alla laadida WinoGrande projekti veebilehelt (Allen Institute for Artificial Intelligence and University of Washington 2025). Andmestik on koostatud JSONL-vormingus (*JavaScript Object Notion Lines format*), mis ei ole tõlkeabi- ja tekstitöötlusprogrammide jaoks vastuvõetav vorming. Iga ülesande õiget vastusevarianti tähistab ülesande qID-koodi (joonis 1) viimane number.


```
1  {"qID": "3QX22DUVOQVY79AVVUY9TVTO8KCVM0-2", "sentence": "Kenneth went cheap on the
   gemstone present for Michael and _ was understanding about being a cheapskate.",
   "option1": "Kenneth", "option2": "Michael"}
2  {"qID": "3LOJFQ4BOZTHN4A6JXLZ8WVDJI4DKP-2", "sentence": "There were more holes in the yard
   of Amy than Carrie because _ had less dogs at their home.", "option1": "Amy", "option2":
   "Carrie"}
3  {"qID": "3INZSNUD824X68RFF0UAINHY9KB9D5-2", "sentence": "The dog didn't like its collar
   but was okay with its leash because the _ was loose on it.", "option1": "collar",
   "option2": "leash"}
```


Joonis 1. WinoGrande andmestik JSONL-vormingus

JSONL-vormingus andmestik teisendati vabavaralist JSON-konverterit (JSON Formatter &Validator 2025) kasutades tõlkeabiprogrammide jaoks vastuvõetavasse JSON-vormingusse (*JavaScript Object Notations format*). Käesolevas magistritöös kasutati tõlkeprotsessi hõlbustamiseks tõlkeabiprogrammi The Phrase Localization Platform (The Phrase Localization Platform 2025). Tõlkeabiprogrammi kasutati eelkõige töövoo haldamiseks. Tõlkeabiprogrammi sisseehitatud masintõlkefunktsioon küll toetas tõlkeprotsessi, kuid väljundit tuli järeltoimetamise käigus erinevatel põhjustel oluliselt muuta.

Tõlgitud andmestik asub järgneval lingil: https://huggingface.co/datasets/tartuNLP/winogrande_et

Sellel lingil asuv andmestik ülesannete õigeid vastuseid ei sisalda, kuna vastuste kättesaadavaks tegemine veebis võib lühendada andmestiku praktilist eluiga. Käesolevas magistritöös on näitena esitatud ülesannete õige vastusevariant märgitud paksus kirjas (vt nt tabel 4).

2.2 Võrdlemine masintõlkega

Varasemalt on näidatud, et testandmestike masintõlkimine mõjutab andmestike kvaliteeti ja sellest tulenevalt ka keelemudelite testisooritust (Plaza jt 2024). Samuti on kirjeldatud sihtkeele omapäradest ja tõlkevastete mitmetähenduslikkusest tulenevaid probleeme, mille tõttu võivad ülesanded masintõlgitud kujul oma eesmärgi kaotada (Emelin ja Sennrich 2021; Vádasz ja Ligeti-Nagy 2022). Kuna eesti keele puhul ei ole varem analüüsitud, kas WinoGrande andmestikku oleks võimalik masintõlgitud kujul eesti keeles kasutada, esitatakse selle magistritöö tõlkeprobleemide analüüsis võrdluseks ka analüüsis käsitletavate Winogradi ülesannete masintõlgitud versioon. Masintõlge on valminud Tartu Ülikooli arvutiteaduse instituudis ning see on tehtud keelemudeliga Open AI GPT-4. Tõlkimiseks kasutatud viip on esitatud käesoleva töö lisas 1. Võrdlusena esitatud masintõlget tõlkeprotsessis ei kasutatud. Masintõlke kaasamise eesmärk on hinnata, kas selle ja taoliste andmestike tõlkimine eesti keelde ainuüksi masintõlkega võiks olla sobilik lahendus. Keelemudeleid testitakse lisaks käesoleva magistritöö käigus valminud tõlgitud andmestikule ka masintõlgitud andmestikuga, et hinnata tõlke kvaliteedi mõju testisooritusele kvantitatiivselt.

2.3 Tõlkeprobleemide analüüs

2.3.1 Käänded

WinoGrande ingliskeelses andmestikus esinevad ülesande mõlemad vastusevariandid ülesandes muutmata kujul. Näiteks tabelis 4 toodud ülesandes on vastusevariandid *chatroom* ja *classroom* ülesandes esitatud samal kujul. Erandiks on omastavas käändes nimed, mida esineb WinoGrande ingliskeelse andmestiku 87-s ülesandes (vt nt tabel 8). Eestikeelses tõlkes ei ole võimalik sarnast ülesehitust säilitada, sest eesti keeles võivad vastusevariandid ja nendele vastavad nimed, nimisõnad või nimisõnafraasid ülesandes esineda erinevates käänetes. Eestikeelses tõlkes on tabelis 4 toodud ülesande vastusevariandid omastavas käändes ning ülesandes nendele vastavad nimisõnad on seesütlevas käändes.

Tabel 4

Käänete esinemine lähte- ja sihttekstis

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*The man could learn English in a classroom or a chatroom, and chose the _ since it avoided commitment.*	***chatroom***	*classroom*
Masintõlge	Mees võis õppida inglise keelt klassiruumis või jututoas, ja valis _ kuna see vältis pühendumist.	**jututoa**	klassiruumi
Tõlge	Mees võis inglise keelt õppida klassiruumis või jututoas, ja ta valis _, kuna see polnud kohustuslik.	**jututoa**	klassiruumi

Tegemist on kahe andmestiku vahelise olulise erinevusega, kuna ingliskeelses andmestikus viitab anafooriline viitesuhe reeglina nime, nimisõna või nimisõnafraasi samale vormile, kuid eestikeelses andmestikus võib viidatav keeleüksus esineda erinevates käändevormides. Magistritöö käigus valminud tõlge sisaldab 1398 ülesannet (79% ülesannetest), milles vastusevariantidest üks või mõlemad esinevad ülesandes vastusevariandist erinevas käändes. Ülejäänud 369-s ülesandes (21% ülesannetest) esinevad vastusevariandid ülesandes muutmata kujul (vt tabel 5).

Kuna tõlgitud andmestik sisaldab näiteid, milles käänded vastusevariantide ja ülesandes esinevate sõnade vahel ühtivad, kui ka näiteid, milles käänded erinevad, siis on andmestiku põhjal võimalik edaspidi analüüsida ka seda, kas käänded tekitavad keelemudelile anafooriliste viitesuhete lahendamisel raskusi.

Masintõlge ei tule vastusevariantide tõlkimisega ühtlaselt toime. Masintõlkes leidub vastusevariante, mis on tõlgitud kontekstist lähtuvalt (vt tabel 4), ning vastusevariante, mis on tõlgitud kontekstist eraldiseisvalt ning pole seega õiges käändes (vt tabel 6).

Tabel 5

Vastusevariandid esinevad tõlgitud ülesandes muutmata kujul

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*James could not believe the bottle is already empty without filling up the cup. The _ is too small.*	***bottle***	*cup*
Masintõlge	James ei suutnud uskuda, et pudel on juba tühi, ilma et ta oleks tassi täitnud. _ on liiga väike.	**pudel**	tass
Tõlge	Jaak ei suutnud uskuda, et pudel on juba tühi ning tops pole ikka veel täis. See _ on liiga väike.	**pudel**	tops

Käänded tekitasid raskusi lausepaaride tõlkimisel, kuna vastavalt ülesannete koostamise aluseks olnud põhimõtetele peab lausetevaheline sõnade kattuvus olema vähemalt 70% (Levesque jt 2021) ja lausepaaril peavad olema samad vastusevariandid (Sakaguchi jt 2019). Tabelis 6 on näitena esitatud üks seesugune lausepaar, milles on nendele tingimustele vastava tõlke koostamine raskendatud. Lausepaari esimese lause tegusõna *puudutama* laiendiks olev sõna peab eesti keeles olema osastavas käändes (puudutama pliiti). Lausepaari teise lause tegusõna *põletama* laiendiks olev sõna peab eesti keeles olema kaasaütlevas käändes (põletama leegiga). Seega tuleb tõlkides lause ümber sõnastada, et vastusevariandid oleksid samas käändes.

Lisaks käänete ühildamisele tuleb tõlkes lahendada ka liitsõnast *gas stove* tulenev tõlkeprobleem. Vastusevariandid *stove* (pliit) ja *gas* (gaas) on liitsõna *gas stove* (gaasipliit*)* moodustajad. Kui tõlgitud ülesandesse jääb sõna *gaasipliit*, siis selles vastusevariante *pliit* ja *gaas* ei esine. Seda tõlkeprobleemi kirjeldasid ka Emelin ja Sennrich (2021), kes otsustasid masintõlgitud andmestikust taolised ülesanded välja jätta. Käesolevas magistritöös tõlgiti kõik WinoGrande testandmestikku kuuluvad ülesanded ning tõlkimisel lähtuti metoodikast, mida kasutasid Amsili ja Seminck (2017), kes asendasid WinoGrande Schema Challenge'i andmestikku tõlkides probleemsed sõnad uute sõnadega, mis võimaldasid ülesande eesmärgi säilitada.

Tabelis 6 toodud ülesande tõlkes on kasutatud sõnakombinatsiooni *pliidil gaasi*, kuna see annab jätkuvalt edasi informatsiooni, et tegemist on gaasipliidiga, kuid paigutab mõlemad

vastusevariandid ülesandesse. Sarnane probleem esineb liitsõnaga *gaasileek*. Tõlkes on kasutatud sõnastust, mis võimaldab kasutada sõnu *gaas* ja *leek* eraldiseisvana ning lauseid on kohandatud, et nende sõnadevaheline kattuvus oleks nõuetekohaselt vähemalt 70%.

Tabel 6

Lausepaaride tõlke 70%-lise kattuvuse ja vastusevariantide samasuse tingimuse täitmine

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande Lausepaari 1. lause	*When I tried to turn off the gas stove, my hand got too close to the burner and I burnt some skin off touching the _.*	***stove***	*gas*
Masintõlge Lausepaari 1. lause	Kui proovisin gaasipliiti välja lülitada, sain oma käe põletile liiga lähedale ja põletasin osa nahka maha, puudutades _.	**pliit**	gaas
Tõlge Lausepaari 1. lause	Kui ma püüdsin pliidil gaasi välja lülitada, läks mu käsi põletile liiga lähedale ja mu nahk sai _ puudutamisel kõrvetada.	**pliidi**	gaasi
WinoGrande Lausepaari 2. lause	*When I tried to turn off the gas stove, my hand got too close to the burner and I burnt some skin off with the _ flames.*	*stove*	***gaas***
Masintõlge Lausepaari 2. lause	Kui proovisin gaasipliiti välja lülitada, sattus mu käsi liiga lähedale põletile ja ma põletasin osa nahka _ leekidega.	pliit	**gaas**
Tõlge Lausepaari 2. lause	Kui ma püüdsin pliidil gaasi välja lülitada, läks mu käsi põletile liiga lähedale ja mu nahk sai _ põlemisel tekkinud leegiga kõrvetada.	pliidi	**gaasi**

2.3.2 Nimede lokaliseerimine

WinoGrande testandmestik sisaldab 1072 Winogradi ülesannet, mille vastus on inimese nimi (vt nt tabel 7). Lisaks on andmestikus inimese nimesid sisaldavaid ülesandeid, mille vastus ei ole nimi (vt nt tabel 9). Kõik andmestikus kasutatud nimed on inglise nimed (nt Mary, Samantha, Michael) ning nimed võivad andmestikus korduda. Andmestiku tõlkimisel lokaliseeriti kõik inglise nimed ja asendati need eesti nimedega (nt Mari, Sandra, Mihkel). Nimed on tõlgitud ülesannetes ja vastusevariantides üldiselt erinevates käändevormides (vt tabel 7), harvemini tuleb ette näiteid, milles nimed on tõlgitud lauses ja vastusevariantides samas käändes (vt tabel 8). Tõlgitud andmestikuga on edaspidi võimalik analüüsida ka seda, kas eesti nimede käändevormid raskendavad anafoorilise viitesuhte lahendamist. Masintõlkes on kõik vastusevariantides esinevad nimed nimetavas käändes sõltumata kontekstist (vt tabel 7).

Nimede lokaliseerimisel järgiti andmestiku koostamise aluseks olnud põhimõtet, mille kohaselt peavad mõlemad ülesandes mainitud inimesed olema sõltumata kontekstist samast soost. Kõik inglise naisenimed lokaliseeriti eesti naisenimedeks ja inglise mehenimed eesti mehenimedeks. Ülesandes mainitud nimed peavad olema samast soost, kuna on näidatud, et keelemudelid võivad anda väljundi, mis sisaldab soolisi eelarvamusi (Zhao jt 2024). Andmestik sisaldas ühte ülesannet, milles nimed on taotluslikult erinevast soost ning see erand säilitati tõlkes (tabel 9).

Tabel 7

Nimed on tõlkes erinevates käändevormides

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*Rebecca prefers to be clean every day unlike Monica, so _ likes to take more showers.*	**Rebecca**	Monica
Masintõlge	Rebecca eelistab olla iga päev puhas erinevalt Monicast, seega _ meeldib talle sagedamini duši all käia.	**Rebecca**	Monica
Tõlge	Reelika eelistab erinevalt Moonikast iga päev puhas olla, seega meeldib _ sagedamini duši all käia.	**Reelikale**	Moonikale

Tabel 8

Nimed on tõlkes samades käändevormides

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*Megan washed Jessica's hat with the load of laundry, so _ felt guilty that the hat was ruined.*	**Megan**	Jessica
Masintõlge	Megan pesi Jessica mütsi koos pesuga, seega _ tundis end süüdi, et müts oli rikutud.	**Megan**	Jessica
Tõlge	Mare pani koos ülejäänud pesuga pesumasinasse ka Jaanika mütsi, seega tundis _ end mütsi rikkumise pärast süüdi.	**Mare**	Jaanika

Tabel 9

Nimed on ülesandes taotluslikult erinevast soost

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*In their marriage, Diane does the dishes and Don does the cooking. They thought _ was women's work.*	***dishes***	*cooking*
Masintõlge	Nende abielus teeb Diane nõusid ja Don teeb süüa. Nad arvasid, et _ oli naiste töö.	**nõusid**	süüa tegemine
Tõlge	Tiina peseb nende abielus nõusid ja Tõnu teeb süüa. Nad leidsid, et _ on naiste töö.	**nõude pesemine**	söögi tegemine

2.3.3 Vastusevariantide järjekord

Andmestiku koostamise aluseks olnud põhimõtetes on märgitud tingimus, mille kohaselt peab iga ülesande esimene vastusevariant esinema lauses esimesena ja teine vastusevariant teisena (Levesque jt 2012). WinoGrande testandmestikus on 218 ülesannet, milles

vastusevariantide järjekord on vastupidine (vt nt tabel 4). Tegemist ei ole keelelise iseärasusega, kuna andmestiku koostamisel oleks olnud võimalik vastusevariandid vastavas järjekorras esitada. Võib oletada, et kuigi WinoGrande andmestiku koostamisel on üldiselt järgitud tingimusi, mida Levesque jt (2012) Winograd Schema Challenge'i koostamiseks kasutasid, siis seda tingimust ei ole WinoGrande andmestiku koostamisel üle võetud. Kuna tõlkes püüti säilitada vastuste üksühene võrreldavus, siis seda anomaaliat ei parandatud. Kui ülesande õige vastus oli esimene vastusevariant, on ka tõlkes ülesande õige vastus esimene vastusevariant, sõltumata sellest, kas esimene vastusevariant esines lauses esimesena.

Tõlkides lisandus ülesandeid, milles vastusevariandid ei esine lauses samas järjekorras, kuna eesti keeles esineb konstruktsioone, mille kasutamine eeldab lauseliikmete ümber tõstmist ülesandes. Näiteks eeldab lauseliikmete ümbertõstmist tagasõna *asemel* kasutamine (tabel 10). Kuigi lauset oleks võimalik lauseliikmete järjekorra säilitamiseks ümber kirjutada, muudaks see laused põhjendamatult pikaks ja kohmakaks. Näiteks tabelis 10 toodud tõlkes saaks moodustada järgneva lause: „Nahaarst valis oma uurimistöö subjektiks Juku, selle asemel et valida Kalev, sest _ polnud aknet."

Tabel 10

Tagasõna *asemel* eeldab lauseliikmete ümbertõstmist ülesandes

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*The dermatologist chose Brett for his research project rather than Kenneth because _ had no acne.*	*Bett*	***Kenneth***
Masintõlge	Dermatoloog valis oma uurimisprojekti jaoks Bretti, mitte Kennethit, sest _ -l ei olnud aknet.	Brett	**Kenneth**
Tõlge	Nahaarst valis oma uurimistöö subjektiks Kalevi asemel Juku, sest _ polnud aknet.	Jukul	**Kalevil**

2.3.4 Lokaliseerimine

Ülesannete tõlkimisel lokaliseeriti laused, mis sisaldasid võõrapäraseid kohanimesid või sõnu. Tabelis 11 esitatud näites on kaktusekommid (inglise keeles *cactus candy*) asendatud sõnaga *sõir* ja kohanimed (Arizona, New Mexico, Sedona, Carlsbad Caverns) on asendatud Eesti kohanimedega (Põlvamaa, Võrumaa, Räpina, Piusa koopad).

Tabel 11

Kohanimede ja võõrapäraste viidete lokaliseerimine

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*I was trying to decide on whether or not to get cactus candy from Arizona or New Mexico.* *Seeing as I was in Sedona I picked _.*	**Arizona**	New Mexico
Masintõlge	Ma üritasin otsustada, kas osta kaktusekommi Arizonast või Uus-Mehhikost. Kuna ma olin Sedonas, valisin _.	**Arizona**	Uus-Mehhiko
Tõlge	Ma püüdsin otsustada, kas osta sõira Põlvamaalt või Võrumaalt. Kuna olin Räpinas, valisin _.	**Põlvamaa**	Võrumaa
WinoGrande	*I was trying to decide on whether or not to get cactus candy from Arizona or New Mexico. Seeing as I was in Carlsbad Caverns I picked _.*	*Arizona*	***New Mexico***
Masintõlge	Ma proovisin otsustada, kas osta kaktusekommi Arizonast või Uus-Mehhikost. Kuna ma olin Carlsbadi koopas, valisin _.	Arizona	**Uus-Mehhiko**
Tõlge	Ma püüdsin otsustada, kas osta sõira Põlvamaalt või Võrumaalt. Kuna olin Piusa koobastes, valisin _.	Põlvamaa	**Võrumaa**

Lokaliseerides kasutati võimalust tuua eestikeelsesse andmestikku sisse näiteks Eesti geograafilisi iseärasusi (tabel 12), eestlastele tuttavaid kaubamärke ja tooteid. Kuna andmestikuga testitakse keelemudeli argimõistuslikku järeldamisoskust eesti keeles, võimaldab see hinnata, kas keelemudel suudab teha seesuguseid järeldusi ka näiteks Eesti geograafia ja kultuuriruumi kohta. Masintõlkega ei ole võimalik seesuguseid andmestikke kindlale kultuuriruumile kohandada.

Tabel 12

Eesti geograafilisi iseärasusi hõlmav lokaliseerimine

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*Samantha wants to go to college in Florida, but Jessica does not; this is due to _ loving warm winters.*	**Samantha**	Jessica
Masintõlge	Samantha tahab minna Floridasse ülikooli, kuid Jessica ei taha; see on tingitud _ armastusest sooja talve vastu.	**Samantha**	Jessica
Tõlge	Sandra tahab Tallinnasse ülikooli minna, kuid Jaanika ei taha. _ meeldib mere ääres elada.	**Sandrale**	Jaanikale

2.3.5 Mitmussõnad

WinoGrande andmestik sisaldab ülesandeid, mille vastusevariandid on inglise keeles ainsuses nimisõnad, kuid millest ühe vastusevariandi eestikeelne tõlkevaste on mitmussõna (tabel 13). Tegemist on omapärase tõlkeprobleemiga, mida on kirjeldanud ka Emelin ja Sennrich (2021), kes täheldasid sama probleemi Winogradi ülesannete prantsuse keelde tõlkimisel. Kui tõlkes on üks vastusevariant ainsuses ja teine mitmuses, ei täida ülesanne oma eesmärki, kuna keelemudel võib valida sobiliku vastusevariandi tegusõna ja nimisõna vorme võrreldes.

Tabelis 13 toodud näites on sõnad *beard* ja *moustache* inglise keeles ainsuse vormis. See tähendab ühtlasi, et selles Winogradi ülesandes esinevale lüngale järgnev tegusõna ainsuse vorm (*was*) ei välista kummagi vastusevariandi sobimist lünka. Keelemudel ei saa selles Winogradi ülesandes ainuüksi nimisõna ja tegusõna vormi põhjal otsustada, kas lünka käib sõna *beard* või

moustache. Inglise keelest eesti keelde tõlkides tekib selles näites lause, kus sõna *beard* on tõlkes jätkuvalt ainsuses (habe), kuid sõna *moustache* on eesti keeles mitmussõna (vuntsid). Seega oleks tõlkes üks vastusevariant ainsuses ja teine mitmuses. Originaalitruu tõlke puhul võib keelemudel valida lünka sobiva sõna nimisõna ja tegusõna vormi võrreldes. Kuna see ei ole Winogradi ülesannete püstitus, tuleb tõlget eesmärgi säilitamiseks toimetada. Eestikeelses tõlkes on sõna *habe* asendatud sõnaga *juuksed*. Mõlemad vastusevariandid on seega samas vormis ning seeläbi on säilitatud ülesande eesmärk.

Masintõlgitud ülesanne ei ole eesmärgipärane, kuna masintõlgitud ülesandes on tegusõna ainsuse vormis (*oli*) ja sellest tulenevalt sobiks lünka vaid ainsuses vastusevariant *habe*. Kuigi vastusevariant on masintõlkes õigesti tõlgitud, võib keelemudel selles näites valida masintõlke põhjal lünka vale vastuse, kuna see on ainus vastus, mis sobib grammatiliselt lausesse.

Tabel 13

Mitmussõnade esinemine sihtkeeles

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*When the boys were getting their hair cut, they got the beard done before the moustache, since the _ was thick.*	*beard*	***moustache***
Masintõlge	Kui poistel lõigati juukseid, siis tehti habe enne vuntse, kuna _ oli paks.	habe	**vuntsid**
Tõlge	Kui poisid juuksuris käisid, said juuksed valmis enne vuntse, sest _ olid paksud.	juuksed	**vuntsid**

2.3.6 Vigade parandamine

WinoGrande testandmestikus esines erinevat tüüpi vigadega ülesandeid. Kuna veaga ülesanne ei täida oma eesmärki, siis tõlkes vead parandati. Vigade hulka kuulusid näiteks trükivead, vigased laused ja ülesandepüstitusele mittevastavad ülesanded. Tabelis 14 toodud näites esineb trükiviga sõnas *door knob.* Trükivea tõttu on moodustunud uus tähenduslik sõna *know*, mis keelemudeliga OpenAI GPT-4 tehtud masintõlkes ei kajastu. On näidatud, et see keelemudel on võimeline seesuguseid vigu ära tundma ja konteksti põhjal parandama (Cao jt 2023).

Tabel 14

WinoGrande andmestiku ülesanne, milles esineb trükiviga

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*Ian got shocked by the metal door know when Robert didn't because _ had rubbed their feet across the carpet causing static electricity.*	***Ian***	*Robert*
Masintõlge	Ian sai metallist ukse käepidemest šoki, kui Robert ei saanud, sest _ oli oma jalgu vaiba vastu hõõrunud, tekitades staatilist elektrit.	**Ian**	Robert
Tõlge	Jaan sai metallist uksenupust särtsu, kuid Raul mitte. _ oli hõõrunud vaibal jalgu, tekitades staatilist elektrit.	**Jaan**	Raul

WinoGrande andmestik sisaldas ka ülesandeid, mille vastuseks sobisid mõlemad vastusevariandid (tabel 15). Kuna seesugused ülesanded ei ole eesmärgipärased, kohandati neid ülesandeid tõlkes, et moodustada eestikeelse andmestiku tarbeks ülesandepüstitusele vastav ülesanne. Kui ülesannete sisu eesmärgi säilitamiseks või taastamiseks muudeti, peeti silmas, et õige vastus oleks endiselt sama vastusevariant. Näiteks kui ülesande õigeks vastuseks oli WinoGrande andmestikus märgitud teine vastusevariant, oli ka tõlkes kohandatud ülesande õige vastus teine vastusevariant.

Lisaks sisaldas WinoGrande andmestik ülesandeid, milles oli õigeks märgitud vale vastusevariant (tabel 16). Ka seesuguste ülesannete puhul oli vaja tõlkes ülesande sisu muuta, kuna vale vastusega ülesanne ei täida oma eesmärki. Tabelis 16 toodud näites on ülesande tõlkes nimed ära vahetatud. See parandus tagab, et eestikeelne andmestik sisaldab ülesandepüstitusele vastavaid ülesandeid.

Tabel 15

WinoGrande andmestiku ülesanne, mille vastuseks sobiksid mõlemad vastusevariandid

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*The fair has many fun things to do, such as the rides or games, but the _ have short lines because they are less exciting.*	*rides*	***games***
Masintõlge	Laadal on palju lõbusaid tegevusi, nagu lõbustuspargid või mängud, kuid _ järjekorrad on lühikesed, sest need on vähem põnevad.	lõbustuspargid	**mängud**
Tõlge	Laadal oli palju põnevaid tegevusi, näiteks atraktsioonid ja mängud. _ järjekorrad on siiski lühemad, sest need nõuavad osavust.	Atraktsioonide	**Mängude**

Tabel 16

WinoGrande andmestiku ülesanne, mille õigeks märgitud vastusevariant on vale

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*Elena installed new carpet for Samantha's home, and _ was very tired from all of the hard work.*	*Elena*	***Samantha***
Masintõlge	Elena paigaldas Samantha koju uue vaiba ja _ oli kogu raske töö tõttu väga väsinud.	Elena	**Samantha**
Tõlge	Sandra paigaldas Helena koju uue vaiba ja _ oli kogu sellest raskest tööst väga väsinud.	Helena	**Sandra**

2.3.7 Lausepaaride kattuvus

Andmestik sisaldab 387 lausepaari vormingus ülesannet. Lausepaaride tõlkimisel on vaja jälgida, et lausepaari sõnade kattuvus oleks vähemalt 70% (Levesque jt 2021) ja nende vastusevariandid oleksid samad (Sakaguchi jt 2019). Lausepaaride sõnade kattuvust kontrolliti arvutuslikul meetodil järgneva valemiga:

$$Kattuvuse\ protsent = (2 * kattuvate\ sõnade\ arv)\ /\ sõnade\ koguarv\ kahes\ lauses\ kokku$$

Tõlkes saavutati täielik vastavus mõlemale eelmainitud tingimusele.

Lisaks sellele oli WinoGrande andmestiku koostamise aluseks olnud tingimustes kirjas, et lausepaaride laused peavad olema 15–30 sõna pikad (Levesque jt 2021). Siinkohal on oluline märkida, et see tingimus on kirja pandud ingliskeelse andmestiku jaoks ning keelte erinevast morfoloogilisest tüpoloogiast tulenevalt ei ole seda otstarbekas eesti keelele samal kujul kohaldada. Eestikeelse tõlgitud andmestiku lausepaaride lausete sõnade arv on 9–30 sõna.

Masintõlge käsitles lauseid eraldiseisvana, mistõttu võivad lausepaarid masintõlkes oluliselt erineda. Tabelis 17 esitatud näites on masintõlge pakkunud sõna *trophy* vasteks lausepaari esimeses lauses *karikas*, aga teises lauses *trofee*. Sellest tulenevalt on ka lausepaari vastusevariandid erinevad. Lisaks sellele on lausepaari ülesande tekst masintõlkes erinevalt tõlgitud. Näiteks on lausepaari esimeses lauses kasutatud tegusõna *soovis* ja lausepaari teises lauses kasutatud tegusõna *tahtis*.

Tabel 17

Masintõlge ei säilita lausepaarides sõnade kattuvust

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande Lausepaari 1. lause	*Ben wanted to win both the money and the trophy. However, he liked the _ best because he was poor.*	***money***	*trophy*
Masintõlge Lausepaari 1. lause	Ben soovis võita nii raha kui ka karika. Kuid talle meeldis kõige rohkem _, sest ta oli vaene.	**raha**	karikas
Tõlge Lausepaari 1. lause	Paul tahtis võita nii raha kui ka karika. Talle meeldis _ siiski rohkem, sest ta oli vaene.	**raha**	karikas
WinoGrande Lausepaari 2. lause	*Ben wanted to win both the money and the trophy. However, he liked the _ best because he was rich.*	*money*	***trophy***
Masintõlge Lausepaari 2. lause	Ben tahtis võita nii raha kui ka trofee. Siiski meeldis talle kõige rohkem _, sest ta oli rikas.	raha	**trofee**
Tõlge Lausepaari 2. lause	Paul tahtis võita nii raha kui ka karika. Talle meeldis _ siiski rohkem, sest ta oli rikas.	raha	**karikas**

2.3.8 *Sest*-põhjuslaused

WinoGrande testandmestik sisaldab üle 800 põhjuslause. Eesti keeles väljendab põhjuslause tüüpjuhul pealausega tähistatud sündmuse põhjust ning põhjuslausete hulka kuulub *sest*-kõrvallause, mis järgneb alati pealausele (Erelt 2017: 718). Tabelis 18 toodud näites ei saa tõlkes

sest-kõrvallauset kasutada, kuna lausest võiks järeldada, et Jenniferile ei meeldinud pildistamine, sest Megan oli oma välimuse suhtes enesekindel. Selliste lausete tõlkimisel moodustati tõlkes kaks eraldi lauset ilma *sest*-korrelaatsidendita. Tegemist on lubatud ülesandestruktuuriga, mida on WinoGrande ingliskeelses testandmestikus kasutatud 317-s ülesandes (vt nt tabel 5). Keelemudel lähtub vastuse tuletamisel endiselt samast informatsioonist, kuid see ei asu kausaalsidendi suhtes süntaktiliselt vales lauseosas.

Tabel 18

sest-kõrvallause asub WinoGrande andmestiku ülesandes vales lauseosas

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande	*Megan liked to have their photo taken but Jennifer did not because _ was very confident about their appearance.*	***Megan***	*Jennifer*
Masintõlge	Meganile meeldis, kui nende foto tehti, kuid Jenniferile ei meeldinud, sest _ oli oma välimuse suhtes väga enesekindel.	**Megan**	Jennifer
Tõlge	Marele meeldis, kui teda pildistati, kuid Jannele see ei meeldinud. _ oli oma välimuse suhtes väga enesekindel.	**Mare**	Janne

Kui seesugune probleem esines vaid ühes lausepaari lauses, tehti tõlkes muudatus mõlemas lausepaari lauses, kuna see võimaldab säilitada lausepaaride ühtse vormingu. Tabelis 19 toodud näites esineb probleem lausepaari esimeses lauses, kuid muudatus on tehtud ka lausepaari teises lauses.

Tabel 19

sest-põhjuslause asub lausepaari esimeses lauses vales lauseosas

	Ülesanne	Vastusevariant 1	Vastusevariant 2
WinoGrande Lausepaari 1. lause	*Patricia had to use a blanket to protect their flowers from frost while Jennifer didn't because _ has a garden.*	**Patricia**	Jennifer
Masintõlge Lausepaari 1. lause	Patricia pidi oma lilli kaitsmiseks kasutama tekki, et kaitsta neid külma eest, samas kui Jennifer ei pidanud seda tegema, sest _ on aed.	**Patricia**	Jennifer
Tõlge Lausepaari 1. lause	Piret pidi oma lilli härmatise eest tekiga kaitsma, aga Janne ei pidanud. _ on aed.	**Piretil**	Jannel
WinoGrande Lausepaari 2. lause	*Patricia had to use a blanket to protect their flowers from frost while Jennifer didn't because _ doesn't have a garden.*	Patricia	**Jennifer**
Masintõlge Lausepaari 2. lause	Patricia pidi kasutama tekki, et kaitsta oma lilli külma eest, samas kui Jennifer ei pidanud seda tegema, sest _ ei ole aeda.	Patricia	**Jennifer**
Tõlge Lausepaari 2. lause	Piret pidi oma lilli härmatise eest tekiga kaitsma, aga Janne ei pidanud. _ pole aeda.	Piretil	**Jannel**

2.4 Keelemudelite hindamise tulemused

Masintõlgitud ja magistritöös tõlgitud andmestikuga hinnati keelemudelite OpenAI GPT-4o, EuroLLM 9B (Martins jt 2024), Llammas (Kuulmets jt 2024), LLama 3.3 70B (Grattafiori jt 2024), LLama 3.1 8B (Grattafiori jt 2024) ja LLama 3.1 405B Instruct (Grattafiori jt 2024) argimõistuslikku järeldamisoskust eesti keeles. Lisaks hinnati samade mudelite sooritust ingliskeelse WinoGrande andmestikuga. Hindamistulemused on esitatud tabelis 20.

Tabel 20

Keelemudelite saavutatud tulemused

Keelemudel	OpenAI GPT-4o	EuroLLM 9B	Llammas	LLama 3.3 70B	LLama 3.1 8B	LLama 3.1 405B Instruct
Masintõlgitud andmestik	79,51%	54,95%	49,97%	70,85%	50,99%	72,95%
Tõlgitud andmestik	83,64%	58,46%	50,37%	73,97%	53,99%	78,78%
Ingliskeelne andmestik	85,06%	52,91%	49,58%	80,31%	56,14%	84,15%

Kõik keelemudelid saavutasid magistritöö raames tõlgitud andmestikuga parema tulemuse kui võrdlusena esitatud masintõlgitud andmestikuga. Kõige parema tulemuse saavutas nii tõlgitud kui ka masintõlgitud andmestikuga keelemudel OpenAI GPT-4o. Kõige madalama tulemuse saavutas mõlema andmestikuga keelemudel Llammas.

Tabelis 21 on näidatud magistritöö raames tõlgitud andmestiku ja masintõlgitud andmestiku tulemuste erinevus. Kõige enam (5,83%) paranes keelemudel LLama 3.1 405B Instruct tulemus ning kõige vähem keelemudel Llammas tulemus (0,4%).

Tabel 21

Tõlgitud andmestiku tulemuste erinevus võrreldes masintõlgitud andmestiku tulemustega

Keelemudel	OpenAI GPT-4o	EuroLLM 9B	Llammas	LLama 3.3 70B	LLama 3.1 8B	LLama 3.1 405B Instruct
Erinevus	+4,13%	+3,51%	+0,4%	+3,12%	+3%	+5,83%

Neli keelemudelit saavutas ingliskeelse andmestikuga parema tulemuse kui eestikeelsete andmestikega (tabelid 22 ja 23). Keelemudelid Llammas ja EuroLLM 9B saavutasid magistritöö raames tõlgitud andmestikuga parema tulemuse kui ingliskeelse andmestikuga. Keelemudeli EuroLLM 9B tulemus oli magistritöö raames tõlgitud andmestikuga 5,55% parem kui ingliskeelse andmestikuga. See oli ka kõige suurem tulemuste erinevus ingliskeelse ja tõlgitud andmestiku vahel.

Tabel 22

Ingliskeelse andmestikuga saadud tulemused võrreldes magistritöö raames tõlgitud andmestikuga saadud tulemustega

Keelemudel	OpenAI GPT-4o	EuroLLM 9B	Llammas	LLama 3.3 70B	LLama 3.1 8B	LLama 3.1 405B Instruct
Erinevus	+1,42%	–5.55%	–0,79%	+6,34%	+2,15%	+5,37%

Masintõlgitud andmestikuga tehtud hindamises oli kõige suurem erinevus keelemudeli LLama 3.1 505B Instruct puhul (tabel 23). See keelemudel saavutas ingliskeelse andmestikuga 11,2% parema tulemuse kui masintõlgitud andmestikuga. Kõige vähem erinesid tulemused keelemudel Llammas puhul, mis saavutas magistritöö raames tõlgitud andmestikuga 0,79% parema tulemuse kui ingliskeelse andmestikuga ning masintõlgitud andmestikuga 0,39% parema tulemuse kui ingliskeelse andmestikuga.

Tabel 23

Ingliskeelse andmestikuga saadud tulemused võrreldes masintõlgitud andmestikuga saadud tulemustega

Keelemudel	OpenAI GPT-4o	EuroLLM 9B	Llammas	LLama 3.3 70B	LLama 3.1 8B	LLama 3.1 405B Instruct
Erinevus	+5,55%	–2,04%	–0,39%	+9,46%	+5.15%	+11,2%

2.5 Järeldused

2.5.1 Tõlkimine

Winogradi ülesandeid on varasemalt erinevatesse keeltesse tõlgitud ning paljud autorid on kirjeldanud konkreetsesse sihtkeelde tõlkimisest tulenevaid tõlkeprobleeme. WinoGrande testandmestiku eesti keelde tõlkimisel esines nii tõlkeprobleeme, mida on teistesse keeltesse tõlkimisel täheldatud, kui ka tõlkeprobleeme, mis on omased eesti keelele.

Amsili ja Seminck (2017), kes tõlkisid Winogradi ülesandeid prantsuse keelde, täheldasid, et otsene tõlge võib ülesande eesmärgi sihtkeeles ära kaotada. Seda tähelepanekut ei saa lugeda tõlkeprobleemiks, vaid tegemist on eelkõige metoodikast tuleneva puudusega. Kuna otsest tõlget ei saa alati kasutada, siis vähendab see tõlkija jaoks masintõlke kasutegurit ning eeldab, et tõlkija leiab loovaid tõlkelahendusi, mis jäävad otsesest tõlkevastest kaugele, kuid võimaldavad ülesannete eesmärki säilitada.

Käesoleva magistritöö käigus tõlgitud ülesannetes tuli teha rohkelt asendusi ja muudatusi, et ülesannete eesmärk säilitada ning mõnel juhul ka luua või taastada. Lisaks sellele on andmestik lokaliseeritud. See tähendab, et tõlgitud ülesanded ei ole alati ingliskeelse andmestiku ülesannetega võrreldavad, kuid samas tagab muudatuste ja paranduste tegemine, et eestikeelset andmestikku saab keelemudelite argimõistusliku järeldamisoskuse hindamiseks eesti keeles iseseisvalt kasutada.

Tõlkimisel tuvastati varasemalt kirjeldatud tõlkeprobleeme, näiteks mitmussõnad ja liitsõnad. Eesti keelde tõlkides kasutati nende lahendamiseks probleemsete sõnade asendamist ning liitsõnade puhul tuli sõnu asendada ja lauseehitust kohandada.

Eesti keelde tõlkimist raskendasid käänded. See tõlkeprobleem on omane just eesti keelele ning see muudab teatud määral ka ülesannete ülesehitust, sest ingliskeelses andmestikus esinevad vastusevariandid ülesandes muutmata kujul, kuid eesti keeles esinevad vastusevariandid ülesandes erinevates käänetes. Seega võib eeldada, et eestikeelsed ülesanded on oma ülesehituselt keelemudelile ka veidi keerukamad. Tõlkides oli kõige raskem leida tõlkelahendusi lausepaaride puhul. Võimalikult originaalitruu tõlke puhul oleks lausepaaride vastusevariandid teatud juhtudel erinevas käändes (vt tabel 6). Selle probleemi lahendamine eeldab loovaid tõlkelahendusi, mis säilitavad lausepaarina esitatud ülesannete vastavuse ülesannete koostamise aluseks olnud põhimõtetele.

Andmestiku tõlkimine hõlmas ka lokaliseerimist. Ülesandeid, mis sisaldasid võõrapäraseid kohanimesid ja sõnu, mida on otstarbekas lokaliseerida, oli andmestiku kogumahtu arvestades vähe. Enamasti oli ülesannetes vaja lokaliseerida inimeste nimesid. Masintõlge käsitleb vastusevariante ülesandest eraldiseisvalt ja sellest tulenevalt on masintõlkes kõik vastusevariantides esitatud nimed nimetavas käändes. Nimed lokaliseeriti tõlkes ning need esitati vastusevariantides õiges käändevormis. Kuna vastusevariandid võivad esineda ülesandes vastusevariandist erinevates käändevormides, võimaldab nimede lokaliseerimine edaspidi analüüsida, kas eestipäraste nimede käändevormide mõistmine võib keelemudelile raskusi tekitada.

WinoGrande andmestik sisaldas hulgaliselt ülesandeid, milles esines erinevat tüüpi vigu. Andmestiku ülesanded on kogutud ühisloome teel ja ülesannetel on erinevad autorid. See selgitab ülesannete kvaliteedi varieeruvust. Tõlkimisel osutus kõige keerulisemaks loogikavigade parandamine. Teatud ülesannete puhul oli raske ülesande taotlust mõista ning see suurendas oluliselt tõlkimisega kaasnevat kognitiivset koormust.

WinoGrande ingliskeelsed ülesanded kaasati andmestikku vaid juhul, kui kaks hindajat kolmest pidasid sama vastust õigeks ja kontrollijad nõustusid, et üks vastusevariant sobib lünka suurema tõenäosusega kui teine (Levesque jt 2021). Seega oleks ka eestikeelse andmestiku puhul sellise kontrolli teostamine edaspidi vajalik, kuna ülesandeid on tõlkes muudetud. See vähendaks võimalikku subjektiivsust, mida pole tõlkides täielikult võimalik välistada.

2.5.2 Masintõlge

Käesolevas magistritöös on võrdlusena esitatud keelemudeliga OpenAI GPT-4 tehtud eestikeelne masintõlge WinoGrande ülesannetest. Masintõlke analüüs näitas, et masintõlkes esinevad süstemaatilised probleemid, mis on eelkõige seotud ülesannete koostamise aluseks olnud põhimõtete järgimisega ning ülesande struktuuriga.

WinoGrande ülesanded sisaldavad lünka ning ülesande vastusevariandid on esitatud ülesandest eraldiseisvana. Masintõlgitud vastusevariandid pole enamasti õiges käändes ja ei sobi seetõttu lünka. Sellel võib olla negatiivne mõju keelemudeli testisooritusele, kuid seda tuleks katsetega täpsemalt hinnata.

Lisaks ei järgi masintõlge ülesannete koostamise aluseks olnud põhimõtteid. See on probleemne eelkõige lausepaaride puhul, mille sõnadevaheline kattuvus peab olema vähemalt 70%

ja mille vastusevariandid peavad olema samad. Kuna masintõlge on tehtud keelemudeliga, oleks seda probleemi võimalik lahendada täpsema viiba koostamisega. Näiteks oleks võimalik lisas 1 esitatud viipa täiendada ja lisada juhis, mis täpsustab, et kui kahe järjestikuse ülesande sõnadevaheline kattuvus on vähemalt 70%, peab nende lausete masintõlge samale kriteeriumile vastama.

Seesuguste kriteeriumite lisamine võiks tagada, et masintõlge vastab ülesande koostamise aluseks olnud põhimõtetele, kuid tõlkija või keelespetsialisti rolli see oluliselt ei lihtsustaks. Kuna sõnade tähendus võib otseses tõlkes teatud määral muutuda, oleks jätkuvalt oluline, et ülesandeid kontrollib spetsialist, kes veendub, et tõlgitud ülesanne vastab ülesandepüstitusele. Näiteks ei tuvasta masintõlge mitmussõnu ega paranda süntaksist tulenevaid loogikavigu. Masintõlkega pole võimalik andmestikku ka lokaliseerida.

2.5.3 Hindamistulemused

Varasemalt on täheldatud, et keelemudelite testisooritust mõjutab tõlke kvaliteet (Thellmann jt 2024). Käesolevas töös esitatud hindamistulemused kinnitavad, et keelemudelite tulemused WinoGrande andmestiku eestikeelse tõlke ja masintõlkega on erinevad. Kõik keelemudelid saavutasid magistritöö raames tõlgitud andmestikuga parema tulemuse kui masintõlgitud andmestikuga. Kõige väiksem erinevus ilmnes keelemudeli Llammas puhul, mis saavutas tõlgitud andmestikuga 0,4% parema tulemuse kui masintõlgitud andmestikuga. Kõige suurem erinevus esines keelemudeli LLama 3.1 405B Instruct puhul, mis saavutas tõlgitud andmestikuga 5,83% parema tulemuse kui masintõlgitud andmestikuga. Kõige suurem tulemuse erinevus ingliskeelse ja masintõlgitud andmestikuga oli 11,2%. See tulemus saadi keelemudeliga LLama 3.1 405B Instruct.

Kuna andmestikus tehti palju parandusi ja muudatusi, ei ole nende võrdlustulemuste põhjal võimalik täpselt öelda, mis tulemusi kõige enam mõjutab. Parema tulemuse võib tagada nii tõlke kvaliteet kui ka algses andmestikus esinevate vigade parandamine. Lisaks esines masintõlgitud andmestikus süstemaatiline probleem vastusevariantide käänetega, mis tõenäoliselt tulemusi mõjutab.

Suurem osa keelemudeleid saavutas ingliskeelse andmestikuga parema tulemuse kui magistritöö raames tõlgitud andmestikuga. Erandiks olid keelemudelid Llammas ja EuroLLM 9B, mis saavutasid tõlgitud andmestikuga parema tulemuse kui ingliskeelse andmestikuga. EuroLLM

9B on keelemudel, mis on loodud Euroopa Liidu keelte jaoks (Martins jt 2024), ning Llammas on keelemudel, mis on spetsiifiliselt eesti keelele kohandatud (Kuulmets jt 2024). See võib selgitada, miks nende keelemudelite tulemused on eestikeelse andmestikuga paremad kui ingliskeelse andmestikuga.

Oluline on märkida ka seda, et kuigi magistritöö raames tõlgitud andmestik vastab üldiselt WinoGrande andmestiku koostamise aluseks olnud põhimõtetele, siis kõiki tingimusi ei ole ilma matemaatilist analüüsi tegemata võimalik täita. Winogradi ülesannete koostamisel on järgitud põhimõtet, mille kohaselt ei või ülesande vastus olla piisavalt suuri keelekorpuseid statistiliselt analüüsides lihtsasti tuletatav. Hetkel ei ole võimalik objektiivselt väita, et see tingimus on eestikeelses andmestikus täielikult täidetud, kuna teatud ülesannete puhul oleks vaja sõnade koosesinemise tõenäosust keelekorpustes statistiliselt analüüsida. Enamik ülesandeid on siiski juba algselt koostatud viisil, mis olenemata keelest välistab vastuse tuletamise korpusanalüüsi statistika põhjal. Näiteks ülesandeid, mille vastus on inimese nimi, ei ole korpuste statistilisele analüüsile toetudes üldiselt võimalik lahendada. Sellised ülesanded moodustavad suurema osa andmestikust ja seega vastab ka suurem osa eestikeelse andmestiku ülesandeid sellele põhimõttele.

WinoGrande ingliskeelse andmestiku ülesandeid on lisaks kontrollitud spetsiifilise algoritmiga, mis välistas ülesanded, milles esines sõnavektorite tasandil teatud kallutatus. Eestikeelset tõlgitud andmestikku pole selle algoritmiga kontrollitud. Võib eeldada, et kuna algsed ülesanded on kontrollitud, siis tõlgitud ülesanded vastavad teatud määral sellele tingimusele. Selleks et tõlge selle põhimõttega kooskõlastada, oleks vaja seda algoritmi rakendada ka eestikeelse tõlgitud andmestiku puhul.

KOKKUVÕTE

Winogradi ülesandeid on mitmesse keelde tõlgitud, kuid nende tõlkimist eesti keelde pole varasemalt kirjeldatud ega analüüsitud. Magistritöö eesmärk oli WinoGrande testandmestiku ülesanded keelemudelite hindamiseks eesti keelde tõlkida ja lokaliseerida ning kirjeldada andmestiku tõlkimisel täheldatud tõlkeprobleeme. Magistritöös analüüsiti ka seda, kas masintõlge võiks olla sobilik lahendus selle ja teiste samalaadsete testandmestike eesti keelde tõlkimiseks.

Käesolevas töös tõlgiti eesti keelde 1767 Winogradi ülesannet. Tõlkimise ja lokaliseerimise tulemusena valmis eestikeelsetest Winogradi ülesannetest koosnev testandmestik, millega hinnati keelemudeleid OpenAI GPT-4o, EuroLLM 9B (Martins jt 2024), Llammas (Kuulmets jt 2024), LLama 3.3 70B (Grattafiori jt 2024), LLama 3.1 8B (Grattafiori jt 2024) ja LLama 3.1 405B Instruct (Grattafiori jt 2024). Andmestiku tõlkimisel ilmnes tõlkeprobleeme, mida olid varasemalt kirjeldanud ka teised autorid, ja tõlkeprobleeme, mis olid omased eelkõige eesti keelele. Eesti keelde tõlkimist raskendasid käänded. Lisaks tõlkimisele tuli teatud hulka ülesandeid erinevatel põhjustel muuta ja kohandada.

Keelemudelite testitulemustest selgus, et kõik keelemudelid saavutasid magistritöö raames tõlgitud andmestikuga parema tulemuse kui masintõlgitud andmestikuga. Keelemudelid OpenAI GPT-4o, LLama 3.3 70B (Grattafiori jt 2024), LLama 3.1 8B (Grattafiori jt 2024) ja LLama 3.1 405B Instruct (Grattafiori jt 2024) saavutasid ingliskeelse andmestikuga parema tulemuse kui eestikeelse andmestikuga. Keelemudelite EuroLLM 9B (Martins jt 2024) ja Llammas (Kuulmets jt 2024) sooritus oli eestikeelse andmestikuga parem kui ingliskeelse andmestikuga.

Magistritöös analüüsiti, kas masintõlge oleks sobilik meetod selle ja taoliste andmestike eesti keelde tõlkimiseks. Keelemudeliga OpenAI GPT-4 valminud masintõlke analüüs näitas, et masintõlge ei pruugi selle andmestiku jaoks sobilik lahendus olla, kuna andmestiku ülesanded peavad vastama etteantud tingimustele, mida masintõlge ei järgi. Kuna keelemudeleid nagu OpenAI GPT-4 saab viipade abil täpsemat väljundit andma suunata, saab käesolevas magistritöös loetletud tõlkeprobleemide lahendusi kasutada täpsemaid juhiseid sisaldavate viipade koostamiseks. Sel viisil oleks sarnaseid andmestikke võimalik edaspidi keelemudelitega tõhusamalt masintõlkida. See hõlbustaks vähesel määral seesuguste andmestike tõlkimist, kuid andmestiku kontrollimine ja kohandamine keelespetsialistide poolt oleks endiselt hädavajalik.

WinoGrande andmestiku ülesanded on koostatud Levesque'i jt (2012) kirjeldatud põhimõtete alusel ning lisaks sellele vastavad WinoGrande ülesanded lisatingimustele, mida on kirjeldanud Sakaguchi jt (2019). Käesoleva magistritöö raames valminud tõlkes on neid põhimõtteid üldiselt järgitud, kuid teatud põhimõtetele vastavust selle töö raames ei kontrollitud, kuna see eeldaks matemaatilist analüüsi. Näiteks on Winogradi ülesannete koostamisel järgitud põhimõtet, mille kohaselt ei või ülesande vastus olla piisavalt suuri keelekorpuseid statistiliselt analüüsides lihtsasti tuletatav. Lisaks sellele on WinoGrande andmestiku ülesandeid kontrollitud spetsiifilise algoritmiga, mis välistas ülesanded, milles esines sõnavektorite tasandil teatud kallutatus. Selleks et WinoGrande andmestiku eestikeelne tõlge oleks kõikide WinoGrande andmestiku koostamise aluseks olnud põhimõtetega täielikult kooskõlas, oleks vajalik edasine töö, mis hõlmaks ka matemaatilisi analüüsimeetodeid.

Magistritöö sisaldab keelemudelite testimise koondtulemusi, kuid andmestiku põhjal on võimalik analüüsida ka üksikute ülesannete või sarnaste ülesannete alamhulkade vastuseid. Edaspidine täpsem analüüs võimaldab tuvastada, mis keelemudelitele eesti keele puhul süstemaatiliselt raskusi tekitab. Tõlgitud andmestikku saab keelemudelite hindamiseks kasutada iseseisva andmestikuna või koos ingliskeelse andmestikuga. Selles töös valminud andmestik jääb veebis kättesaadavaks ning seda kasutatakse eesti keeletehnoloogiate jätkuvaks arendamiseks.

KIRJANDUSE LOETELU

Amsili, P ja Seminck, O. (2017) A Google-Proof Collection of French Winograd Schemas *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes.* Corbon, 24–29

Bandarkar, L., Liang, D., Muller, B., Artetxe, B., Shukla, S. N., Husa, D., Goyal, N., Krishnan, A., Zettlemoyer, L. ja Khabsa, M. (2023) The Belebele Benchmark: a Parallel Reading Comprehension Dataset in 122 Language Variants. *arXiv:2308.16884*

Bernard, T. ja Han, T (2020) Mandarinograd: A Chinese Collection of Winograd Schemas. – *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 21–26

Cao, Q., Kojima, T., Matsuo, Y. ja Iwasawa, Y (2023) Unnatural Error Correction: GPT-4 Can Almost Perfectly Handle Unnatural Scrambled Text *arXiv:2306.01393v3*

Davis, E. (2016) Winograd Schemas and Machine Translation *arXiv:1608.01884*

Eesti Keele Instituudi teatmik (2025). Vaadatud 08.02.2025 https://eki.ee/teatmik/mis-on-suured-keelemudelid/

Emelin, D., Sennrich, R. (2021) Wino-X: Multilingual Winograd Schemas for Commonsense Reasoning and Coreference Resolution *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing.* 8517–8532

Erelt, M. (2017) Liitlause. – M. Erelt (toim) ja H. Metslang (toim). *Eesti keele süntaks*. Tartu: Tartu Ülikooli Kirjastus, 718–719

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A.,

Sravankumar, A., Korenev, A., Hinsvark, A., Rao, A., Zhang, A., Rodriguez, A., Gregerson, A., Spataru, A., Roziere, B., Biron, B., Tang, B., Chern, B., Caucheteux, C., Nayak, C., Bi, C., Marra, C., McConnell, C., Keller, C., Touret, C., Wu, C., Wong, C., Canton Ferrer, C., Nikolaidis, C., Allonsius, D., Song, D., Pintz, D., Livshits, D., Wyatt, D., Esiobu, D., Choudhary, D., Mahajan, D., Garcia-Olano, D., Perino, D., Hupkes, D., Lakomkin, E., AlBadawy, E., Lobanova, E., Dinan, E., Smith, E. M., Radenovic, F., Guzmán, F., Zhang, F., Synnaeve, G., Lee, G., Lewis Anderson, G., Thattai, G., Nail, G., Mialon, G., Pang, G., Cucurell, G., Nguyen, H., Korevaar, H., Xu, H., Touvron, H., Zarov, I., Arrieta Ibarra, I., Kloumann, I., Misra, I., Evtimov, I., Zhang, J., Copet, J., Lee, J., Geffert, J., Vranes, J., Park, J., Mahadeokar, J., Shah, J., van der Linde, J., Billock, J., Hong, J., Lee, J., Fu, J., Chi, J., Huang, J., Liu, J., Wang, J., Yu, J., Bitton, J., Spisak, J., Park, J., Rocca, J., Johnstun, J., Saxe, J., Jia, J. jt (460 lisaautorit) (2024) The Llama 3 Herd of Models *arXiv:2407.21783*

Ivanov, T., ja Penchev, V. (2024) AI Benchmarks and Datasets for LLM Evaluation *arXiv:2412.01020*

Kuulmets, H.-A., Purason, T., Luhtaru, A., Fishel, M. (2024) Teaching Llama a New Language Through Cross-Lingual Knowledge Transfer *Findings of the Association for Computational Linguistics: NAACL 2024* Mexico City, Mexico, 3309–3325

Naveen, P., ja Trojovský, P. (2024) Overview and challenges of machine translation for contextually appropriate translations *iScience* 27 (10), 110878

Longjohn, R., Kelly, M., Singh, S. ja Smyth, P. (2024) Benchmark Data Repositories for Better Benchmarking *arXiv:2410.24100v1*

Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J. G. C., Birch, A., Martins, A. F. T. (2024) EuroLLM: Multilingual Language Models for Europe arXiv:2409.16235

Melo, G.S., Imaizumi, V.A., ja Cozman, F.G. (2019) Winograd Schemas in Portuguese. – *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*. Porto Alegre: SBC, 787–798

Pajusalu, R. (2017) Viiteseosed. – M. Erelt (toim) ja H. Metslang (toim). *Eesti keele süntaks*. Tartu: Tartu Ülikooli Kirjastus, 566–588

Plaza, I., Melero, N., del Plozo, C., Conde, J. ja Reviriego, P. (2024) Spanish and LLM benchmarks: is MMLU lost in translation? *arXiv:2406.17789v1*

Ponti, E. M., Glava, G., Majewska, O., Liu, Q. ja Korhonen, A. (2020) XCOPA:AMultilingual Dataset for Causal Commonsense Reasoning. *arXiv:2005.00333*

Sakaguchi, K. , Le Bras, R., Bhagavatula, C., Choi, Y. (2019) WinoGrande: An Adversarial Winograd Schema Challenge at Scale *arXiv:1907.10641*

Saleh, M. ja Paquelet, S. (2024) Anatomy of Neural Language Models *arXiv:2401.03797*

Sirts, K. (25.10.2024) Eesti keele ja kultuuri toetusest suurtes keelemudelites. *Sirp.* Vaadatud 12.02.2025 https://www.sirp.ee/s1-artiklid/c21-teadus/eesti-keele-ja-kultuuri-toetusest-suurtes-keelemudelites/

Team of researchers to teach Estonian language, culture to language models. *ERR News*, 04.09.2024 kl 16.04 https://news.err.ee/1609443533/team-of-researchers-to-teach-estonian-language-culture-to-language-models

Thellmann, K., Fromm, M., Stadler, B., Buschhoff, J. S., Jude, A., Leveling, J., Flores-Herr, N., Köhler, J., Jäkel, R. ja Ali, M. (2024) Towards Multilingual LLM Evaluation for European Languages. *arXiv:2410.08928*

Vádasz, N. ja Ligeti-Nagy, N. (2022) Winograd Schemata and Other Datasets for Anaphora resolution in Hungarian. *Acta Linguistica Academica*, 69 (4), 564–580

SUMMARY

The aim of this thesis was to translate and localise the WinoGrande benchmarking dataset into Estonian. The thesis documents translation challenges encountered during the process and evaluates the effectiveness of machine translation for adapting similar benchmarks to Estonian.

The main translation challenges identified align with issues that have been reported previously for Winograd schema translations to various languages. Challenges specific to the Estonian language are to do with the case system, which complicates maintaining the benchmark structure.

The analysis of the machine translation generated with OpenAI GPT-4 revealed systematic issues related to the schema structure which includes gap tokens. Whilst improved prompts can mitigate some problems, language specialists remain essential for verifying that translated schemas preserve their intended purpose and functionality in the translation.

The translated benchmark was used to evaluate large language models OpenAI GPT-4o, EuroLLM 9B, Llammas, LLama 3.3 70B, LLama 3.1 8B, and LLama 3.1 405B Instruct. The thesis presents these results alongside comparative performance on the original English benchmark and machine-translated versions.

Performance analysis showed that all tested models achieved better results with the human-translated benchmark compared to the machine-translated version. Whilst OpenAI GPT-4o and the LLama models performed better on the English benchmark than the Estonian version, EuroLLM 9B and Llammas performed better on the Estonian benchmark.

Whilst additional work is needed to ensure the Estonian schemas fully meet WinoGrande benchmark's technical requirements, the initial results show a significant difference between results obtained with the human-translated dataset and machine-translated dataset. The translated dataset will be used to further study the specific challenges LLMs face in processing the Estonian language, and it contributes to the development of Estonian language technologies.

LISAD

LISA 1

Masintõlke viip

```
'\nPlease translate the given example in JSON format into Estonian. Retain the
JSON structure. Do not translate the keys. Translate the values only.\nPlease
output only the JSON object, nothing else.\nExample:\n{"sentence": "Kenneth
went cheap on the gemstone present for Michael and _ was understanding about
being a cheapskate.", "option1": "Kenneth", "option2": "Michael", "answer":
""}\n'
```

Lõputöö autori kinnitus

Olen lõputöö kirjutanud iseseisvalt. Kõigile töös kasutatud teiste autorite töödele, põhimõttelistele seisukohtadele ning muudest allikatest pärinevatele andmetele on viidatud.

Autor: Marii Ojastu ……..

(allkiri)

14.05.2025

(kuupäev)

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Marii Ojastu ,

annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose

WinoGrande andmestiku tõlkimine suurte keelemudelite argimõistusliku järeldamisoskuse hindamiseks eesti keeles ,

(lõputöö pealkiri)

mille juhendaja(d) on Kairit Sirts ja Marika Borovikova ,

reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada Tartu Ülikooli digitaalarhiivi kuni autoriõiguse kehtivuse lõppemiseni;

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;

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

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

Marii Ojastu

14.05.2025