Deep Neural Nets and the Language Instinct

Kristoffer Bakke Tvedt

Csaba Veres

Masters thesis
Department of Information Science and Media Studies
University of Bergen

November 30, 2023

Acknowledgements

I would like to thank Csaba Veres for being my supervisor and mentor during the prolonged ordeal of writing this masters thesis. I would not be able to finish my studies without his insight and motivation. I would also extend my thanks to Kjersti Birkeland Daae and Stephan Kral for making this template available on overleaf, as well as Tore Birkeland and Raymond Nepstad for creating the original version of the template.

My appreciation also goes to my family for motivating and believing in me, even when I decided to delay my studies by another semester.

As a final acknowledgement, I would like to thank my cat, Bollepus, for keeping me company, and my lap warm, while reading numerous articles on machine learning models and universal grammar.

Kristoffer Bakke Tvedt
Bergen, 30.11.2023

Abstract

With modern machine learning models like NLLB200 and Multilingual BERT showing surprisingly good cross-lingual performance on natural language processing tasks, and research showing that training models on high resource languages help with scores when tested on low resource languages with limited data available (*Wu and Dredze* (2020)). Although universal grammar is a hotly debated issue with many in the ML community arguing against it, we believe that this is only possible because of universal grammar. Research done on brain activity during experiments with test subjects learning real grammatical rules and universal grammar-defying rules indicate that a certain part of our brain is active while learning real grammatical rules, while this part of the brain is not active while learning grammatical rules that deviate from the principles of universal grammar (*Musso et al.* (2003)), signifying the existence of universal grammar.

In this masters thesis we will conduct a series of translation experiments with sentences that deviate from the principles of universal grammar on human participants and the machine learning models NLLB200 and ChatGPT. Trough this research, we find that while human participants perform well on these sentences, machine learning models have a tougher time translating these sentences. With these finding, we find some evidence that support our hypothesis.

Contents

Chapter 1

Introduction

1.1 Motivation

Modern transformer based machine learning models can be trained on high resource languages, and the knowledge they acquire can be used for natural language tasks in languages they were not trained on. With this work I want to show that this is possible because of grammar universals.

The general view of machine learning practitioners, however is that there is no such thing as grammar universals, and in *Evans and Levinson* (2009) paper *The Myth of Language Universals: Language Diversity and Its Importance for Cognitive Science*, the authors argue that grammar universals are a myth. In their paper they claim that languages differ so much from one another in terms of sound, grammar, lexicon and meaning that its hard to find any shared structural property. They go on to argue that universal grammar is empirically false, unfeasible or misleading and that rather than referring to universals they refer to tendencies. They claim that the idea of grammar uniformity attributes from most linguists and cognitive scientists only speak European languages which have a lot of similarities in structure, and also that the topic of grammar generalization has split the views of linguistics with no shared rules of argumentation to resolve the issue.

On the other hand, in the paper *Brocas area and the language instinct* by *Musso et al.* (2003) they find that a certain part of our brain called Brocas area had an increase of activation over time when testing with real grammatical rules that follows the principle of universal grammar. Alongside the experiments done with these real grammatical rules, they also tested subjects on constructed grammatical rules where they would take the same lexicon used for Japanese and Italian, but the rules they made were linguistically illegal as they did not adhere to the principles of universal grammar. They found that there was an increase in blood oxygen-level dependent, or BOLD for short, signal for the real language tasks, but not for the constructed ones. Their findings that the activation of Brocas area is independent of the language of subjects suggests a universal syntactic specialization of this area among real languages.

1.2 Problem Statement

The theory of universal grammar (UG) states that a human raised under normal conditions without extreme sensory deprivation will always develop certain properties within a language, like being able to differentiate between nouns and verbs. Although machine learning practitioners deny the existence of universal grammar, there are models that are trained on high resource languages that can use the knowledge they acquire on task for languages they were not trained on. To find proof of universal grammar we will conduct experiments with machine learning models, where we test the pretrained models on sentences that do not adhere to the principles of universal grammar.

1.3 Objectives

Modern transformer based machine learning models can be trained on high resource languages, and the knowledge they acquire can be used for natural language tasks in languages they were not trained on. Our Hypotheses is that this is only made possible because of grammar universals. We expect the machine learning models to perform significantly worse on translations tasks with sentences that are not applying to the principles of universal grammar than on sentences that adhere to these principles. The object of this thesis is to find evidence for universal grammar by doing machine learning experiments, with the hypothesis that machine translation in low resource languages is made possible by regularities inherent in universal grammar.

1.4 Hypothesis

Our hypothesis is that large language models (LLM) are able to learn about low resource languages when trained on high resource languages because they exploit statistical regularities generated by grammars which conform with universal grammar. This makes the prediction that LLMs should have little knowledge of low resource languages which do not have grammars compliant with UG.

- LLMs will perform poorly on the non UG sentences
- LLMs will perform more poorly on non UG sentences which disrupt statistical regularities in the surface form more severely than non UG sentences which disrupt statistical regularities less severely
- Machine learning models trained on next word prediction will be more robust than models trained on translation

1.5 Contribution

Generalization and universality is one of the biggest goals in machine learning research and is considered by many to be an important step towards artificial general intelligence. With this thesis we aim to prove that modern machine learning models perform well on low resource languages because of the universal grammar it learns from high resource languages during training.

1.6 Thesis outline

In this masters thesis we will summarize the work of Musso et al. from their paper on Brocas area and the effect constructed grammatical rules has on our brains. We will also introduce two cutting edge natural language processing (NLP) models in Meta AIs NLLB200 and OpenAIs ChatGPT. The methods chapter of the thesis will outline the methods used for formulating new grammatical rules that deviate from the principles of universal grammar, how we will be testing NLLB200 and ChatGPT on the sentences generated with said rules, as well as how we will test the sentences on human participants. An introduction to how we will score the translations the machine learning models and humans make will conclude the methods section. The output of the experiments will be reviewed in the results section, before concluding the thesis with discussion of what the results mean and future work.

Chapter 2

Background

2.1 Broca's area

The ability to learn a language is unique to humans, and it is considered general knowledge that children learn their mental grammar spontaneously from the way their parents speak. In 1957 Noam Chomsky proposed that for generalizing from a sample of sentences to language as a whole, an innate set of mental computations is required (Chomsky, 1957 as cited in *Musso et al.* (2003)). Chomsky also argued on the basis of linguistic analyses of sentence structure that universal grammar underlies the Babel of languages (Chomsky, 1986, as cited in *Musso et al.* (2003)). In Musso et al. 's paper *Brocas area and the language instinct* they investigate the system underlying the acquisition of learning new linguistic competence of two different languages: Japanese and Italian. The idea behind this research is that when learning a new language like Italian or Japanese or any other real language based on the principles of universal grammar, a certain part of the brain should be active. When learning an artificial language that does not follow universal grammar structure, a different part of the brain should be used. Musso et al. mentions that some claim that nonspecific learning mechanisms which are not associated with particular cognitive domains underlie the acquisition of new linguistic competence (*Musso et al.* (2003)). If this view is correct, that would indicate that the same brain systems should be used when learning both real and unreal languages.

For their first study they had 12 native German speakers who had never been exposed to Italian and had them learn three grammatical rules of Italian as well as three artificial rules of an unreal language made by manipulating the Italian language. For their second study another 11 native German speakers did almost the same thing, but with Japanese. The research team used fMRI to monitor the difference in brain activity between acquisition of real and unreal grammar. Musso et al. analyzed data from eight subjects from each of the tests, and the results show that there was equally high accuracy in both real and unreal language tasks. They also measured the reaction times of the test subjects, and the results show a significant reduction in reaction time, as well as improvement in accuracy over the course of the sessions.

When it comes to the fMRI data, the results show that the effect of performing the classification task showed activation in many different areas of the brain including prefrontal, parietal, anterior cingulate, occipital cortex, inferior and middle temporal gyrus and the cerebellum on both hemispheres compared to the baseline task of staring at a

black screen *Musso et al.* (2003). While the researchers did not find any specific pattern of brain activation for unreal grammatical acquisition for either Japanese or Italian, they did find that an interaction between real and unreal grammatical acquisition was evident in two different parts of the pars triangularis of the left inferior frontal gyrus named Brocas area. The concluding results from the tests was that there was a significant positive correlation between blood oxygen-level dependent (BOLD) signal and accuracy with the tasks based on real language, while there was a significant negative correlation between parameter estimates and learning unreal rules (*Musso et al.* (2003)).

In the discussion section of their research paper, *Musso et al.* (2003) write that the results of their testing show that a significant correlation between the increase in BOLD signal in the left inferior frontal gyrus and the on-line performance for the real language learning tasks, but not for the unreal language tasks. They continue to say that this discovery stands as neurophysiological evidence that learning a new, real, language involves a brain system that is different from that of learning new linguistic competence for an unreal language that breaks the principles of universal grammar (*Musso et al.* (2003)). Another interesting point the researchers bring up in their paper is that the behavioral analysis of the response times show that the test subjects were answering progressively faster when working on real grammatical tasks than when working on tasks with unreal grammar. They speculate that this can be because of proceduralization of rule-knowledge while working on real grammatical tasks, and that the progressive consolidation of knowledge could be passed on to other sentence material. According to Musso and her peers, an indisputable and essential function of Broca's area is the processing of syntactic aspects of language, and activation of this area of the brain is independent of the language of subjects, which according to the researchers suggests a universal syntactic specialization of this area among real languages. Based on their results from this research paper, as well as results from previous work, they posit that Broca's area is specialized for the acquisition and processing of hierarchical, rather than linear, structures, which represent the common character of every known grammar. They concludes that their results indicate that the left inferior frontal gyrus is centrally involved in learning new languages, but only if the language is based on the principles of universal grammar (*Musso et al.* (2003)).

2.2 NLLB200

A few short years ago, high-quality machine translation worked in only a handful of languages. With NLLB-200, we are closer to one day having systems that enable people to communicate with whomever they choose. It's exciting to see what this unlocks in the present and what it could mean for the future as Meta AI continues to push the boundaries of machine translations.

The significance of language in shaping culture, identity, and acting as a lifeline to the world cannot be overstated. However, the unfortunate reality is that a considerable number of languages lack high-quality translation tools, depriving billions of people of access to digital content and hindering their full participation in online conversations and communities in their preferred or native languages. This gap is particularly pronounced for speakers of languages in Africa and Asia, where the absence of adequate

translation tools is acutely felt by hundreds of millions (*Meta* (2022)).

In response to this challenge and with the vision of fostering better connections today and preparing for the metaverse of tomorrow, researchers at Meta AI initiated the No Language Left Behind (NLLB) project. The primary objective of NLLB is to develop advanced machine translation capabilities encompassing a vast array of languages globally. The latest breakthrough in this endeavor is the creation of a singular AI model known as NLLB-200, proficient in translating 200 different languages with state-of-the-art accuracy. Importantly, many of these languages, including Kamba and Lao, were previously underserved or not supported at all by existing translation tools. Notably, NLLB-200 addresses the significant language gap by providing high-quality translations for 55 African languages, a substantial improvement over the limited support for fewer than 25 African languages offered by current widely used tools. Evaluation through the BLEU scores on the FLORES-101 benchmark indicates an average improvement of 44 percent compared to the previous state of the art, with some African and Indian languages experiencing an increase exceeding 70 percent over recent translation systems (*Meta* (2022)).

In a commitment to openness and inclusivity, Meta AI is taking a groundbreaking step by open-sourcing the NLLB-200 model. Furthermore, the research tools utilized in this project are being made available to enable other researchers to expand this work to additional languages, fostering the development of more inclusive technologies. To catalyze real-world applications of NLLB-200, Meta AI is offering grants of up to 200,000 USD to nonprofit organizations (*Meta* (2022)).

The impact of NLLB extends beyond the realms of research and development, with practical applications influencing the translation services provided daily on Meta's platforms, including Facebook News Feed and Instagram. Imagine the scenario of effortlessly comprehending a post in Igbo or Luganda within a favorite Facebook group with just a click of a button. Beyond convenience, highly accurate translations in numerous languages play a pivotal role in identifying harmful content, preventing misinformation, safeguarding election integrity, and combating instances of online exploitation and human trafficking. The expertise and insights garnered from the NLLB project are now being applied to enhance translation systems utilized by Wikipedia editors, exemplifying the widespread influence of this research (*Meta* (2022)).

Addressing the significant disparities in content across various language versions of Wikipedia, Meta AI has partnered with the Wikimedia Foundation to improve translation systems on the platform. Wikipedia, with versions in over 300 languages, faces challenges, especially in languages spoken outside of Europe and North America, resulting in a substantial discrepancy in article numbers. By integrating the technology behind NLLB-200 into the Wikimedia Foundations Content Translation Tool, Wikipedia editors are now equipped to translate articles in over 20 low-resource languages, including 10 languages that were previously unsupported by any machine translation tools on the platform (*Meta* (2022)).

The development of a unified model for hundreds of languages introduces formidable challenges. Unlike well-represented language pairs, many languages lack substantial parallel sentence data, making traditional methods of training translation models challenging. Current models resort to web-mined data, resulting in poor-quality translations due to variations in source text and issues such as incorrect spellings and missing accent marks. Additionally, optimizing a single model for hundreds of languages with-

out compromising performance or quality poses a significant hurdle. While traditional models opt for separate models for each language direction to achieve optimal quality, scaling this approach becomes challenging as more languages are added (*Meta* (2022)).

To overcome these challenges, Meta AI has made strides in architecture, data sourcing, benchmarking, and more. Advances in LASER, their toolkit for zero-shot transfer in natural language processing, have been instrumental in collecting accurate parallel texts in more languages. LASER3, the latest version, utilizes a Transformer model trained in a self-supervised manner with a masked language modeling objective, enhancing performance through a teacher-student training procedure and language-group specific encoding (*Meta* (2022)).

The progression from the 100-language M2M-100 translation model in 2020 to the current NLLB-200 model signifies Meta AI's commitment to overcoming linguistic barriers and fostering a more inclusive digital landscape. As the NLLB project continues to evolve, it not only holds promise for providing enhanced access to digital content but also for facilitating cross-language contributions and information sharing. While challenges persist, the progress made thus far serves as a testament to Meta AI's dedication to fulfilling its mission in advancing the impact of AI on people's everyday lives (*Meta* (2022)).

MetaAI created a process to automatically conjure translation training data for hundreds of languages. The conventional methods employed in the past for training translation models faced challenges when applied to low-resource settings. These scenarios involve situations where there is a scarcity of data for a particular language, encompassing both aligned textual data (bitext, or pairs of translated sentences) and single-language data (monolingual, or data in one language only). It's noteworthy that numerous low-resource languages relied heavily on small, focused bitext datasets, such as excerpts from the Christian Bible, which posed limitations in terms of domain diversity (*NLLB-Team et al.* (2022)).

To provide some background, the accessibility of publicly available bitext data has historically been limited. The strategy Meta AI employed focused on expanding existing datasets by gathering non-aligned monolingual data. They utilized large-scale data mining techniques to identify sentences with a high likelihood of being translations of each other in diverse languages. To make this applicable to a wide array of languages, the initial step involved developing language identification systems (LID) to categorize the language of a given piece of text. Following this, Meta AI meticulously curated the available monolingual data, implementing processes like sentence splitting and LID, along with various filtering mechanisms. Subsequently, the team proceeded with the mining of aligned pairs (*NLLB-Team et al.* (2022)).

The NLLB team discusses the evaluation of Language Identification (LID) models, primarily focusing on the Flores-200 dataset. It compares the performance of the LID model to other open-source models, namely CLD3, LangId, and LangDetect, emphasizing the challenges of working with low-resource languages. The LID model demonstrates superior performance across various language intersections, particularly outperforming other models on the Flores-200 dataset. However, the discussion acknowledges potential issues when applying the model to noisy web data due to language mixing, script variations, and leetspeak. Human evaluation is introduced to address performance gaps, and the text highlights challenges in identifying confusable language pairs. Additionally, it notes the impact of sentence length on prediction ro-

bustness and suggests potential mitigation strategies (*NLLB-Team et al.* (2022)).

The next step in their paper outlines the importance of monolingual data for downstream tasks like bitext mining and language model training, emphasizing the need for high-quality and clean data. The process involves utilizing web data from CommonCrawl and ParaCrawl, applying language identification to convert paragraphs into sentences, and employing heuristics for data cleaning. Analysis of distribution patterns in monolingual data scores guides the determination of detection thresholds. Various heuristics, including length and content ratios, are applied for further data cleaning. The process also involves deduplication, language model filtering, and addressing computational challenges related to processing a large volume of data. The resulting monolingual dataset comprises billions of sentences, with a focus on maintaining high-quality content for subsequent tasks (*NLLB-Team et al.* (2022)).

The final part about training the model outlines an approach to enhance machine translation in low-resource languages, an area often hindered by limited training data. In machine translation, the quality typically improves with the volume of high-quality training data. However, for low-resource languages, parallel corpora are often drawn from specific sources like the Bible or multinational publications, resulting in limited quantity and domain relevance. The focus here is on creating translation training datasets through bitext mining, primarily paired with English but also exploring mining through other language pairs. Bitext mining involves learning a multilingual sentence embedding space and using a similarity measure to determine if two sentences are parallel. The approach uses global mining, comparing all possible pairs in two collections of monolingual texts.Scaling this representation to 200 languages poses challenges, including ensuring comprehensive language learning and addressing imbalances in training data. Training a multilingual sentence encoder for each new set of languages is computationally expensive. To overcome this, a teacher-student distillation technique is applied, training smaller mutually compatible sentence encoders (*NLLB-Team et al.* (2022)).

Related work in mining methodology is discussed, ranging from early approaches like STRAND algorithm to recent advancements leveraging representation learning. Multilingual sentence representation learning methods such as mBERT, XLM, and LASER are explored, with a focus on addressing limitations and improving performance (*NLLB-Team et al.* (2022)).

The student-teacher mining approach involves adapting a massively multilingual sentence encoder teacher model to various low-resource student models. LASER2 serves as the teacher, and LASER3, an improved version, is used as the student model. The teacher-student training is performed on 16 GPUs with specific parameters (*NLLB-Team et al.* (2022)).

A proxy metric, xsim, is introduced for evaluating new encoders, measuring mining-based multilingual similarity search error rate. An end-to-end encoder evaluation is conducted, identifying the best sentence encoder for each language based on xsim scores, followed by mining, adding mined data to existing bitexts, and training a bilingual NMT system. Language-specific encoder training is discussed, showcasing improvements over LASER for various languages and language families, including European minority languages, Creole languages, Berber languages, Malayo-Polynesian languages, and African languages. The text emphasizes the importance of available monolingual data for successful bitext mining (*NLLB-Team et al.* (2022)).

In conclusion, the approach aims to enhance translation quality for 200 languages, primarily relying on bitext mining to create training datasets for low-resource languages, with a focus on scalability, performance evaluation, and ethical considerations (*NLLB-Team et al.* (2022)).

2.3 ChatGPT

In the ever-evolving landscape of artificial intelligence, natural language processing has emerged as a critical domain, driving innovations that redefine our interaction with machines. At the forefront of this transformative journey stands ChatGPT, a revolutionary language model developed by OpenAI. While initially conceived as a versatile conversational agent, ChatGPT's prowess extends far beyond mere chat-based interactions. This introduction delves into the intricacies of ChatGPT, exploring its evolution, architecture, and, more specifically, its role as an increasingly influential translation model.

The inception of ChatGPT can be traced back to the earlier iterations of the Generative Pre-trained Transformer (GPT) series. As part of OpenAI's commitment to advancing the frontiers of language understanding, GPT-3.5 emerged as a flagship model, boasting an unprecedented 175 billion parameters. This colossal scale represents a paradigm shift in NLP, empowering ChatGPT with an unparalleled ability to grasp linguistic nuances and contextual intricacies. At the heart of ChatGPT lies the GPT-3.5 architecture, a testament to the transformative potential of large-scale language models. Built upon the Transformer architecture, GPT-3.5 leverages self-attention mechanisms to analyze and understand complex patterns within vast amounts of text data. This architecture endows ChatGPT with the capacity to generate coherent and contextually relevant text, making it a formidable force in the realm of language understanding and generation(*Techvify-Software* (2023)).

While ChatGPT was initially designed as a conversational agent, its versatility quickly became evident as researchers and developers explored its potential across diverse applications. Among the myriad tasks that ChatGPT can adeptly handle, language translation has emerged as a standout capability. Traditional approaches to translation often involve specialized models, finely tuned for specific language pairs. ChatGPT, with its adaptive and expansive understanding of language, challenges this conventional wisdom, presenting a compelling alternative.

Historically, translation models were developed using rule-based systems or statistical machine translation approaches. These models relied on predefined linguistic rules or parallel corpora for training. ChatGPT, however, signifies a shift in this paradigm. Its unsupervised learning approach, coupled with the immense scale of the GPT-3.5 architecture, allows it to transcend the limitations of traditional models. The model's adaptability to informal language and diverse language pairs positions it as a dynamic and versatile translation tool(*Techvify-Software* (2023)).

Some strengths with ChatGPT include:

- Contextual Understanding: At the core of ChatGPT's translation capabilities lies its unparalleled contextual understanding. The model excels in capturing the nuances of language, ensuring translations that go beyond literal conversions.

- Adaptability: Unlike traditional translation models, which may struggle with informal language or varied communication styles, ChatGPT adapts seamlessly. Its adaptability to different language pairs and communication nuances enhances its utility in real-world scenarios.
- Zero-Shot Translation: An intriguing feature of ChatGPT is its zero-shot learning capability. Without specific training for a particular language pair, the model can perform translations, demonstrating its flexibility in handling diverse linguistic tasks (*Techvify-Software* (2023)).

And some limitations:

- Lack of Specialization: While ChatGPT demonstrates remarkable proficiency, it lacks the specialized training that dedicated translation models undergo. This can lead to variations in translation quality, particularly for intricate or highly specialized domains.
- Sensitivity to Input Phrasing: The model's responses may exhibit sensitivity to subtle changes in input phrasing, introducing variations in translations. This highlights the importance of precise input formulation for optimal results.
- Finite Context Window: Despite its vast parameter count, ChatGPT operates within a finite context window. Extremely long sentences or documents may pose challenges for maintaining context throughout the entire text (*Techvify-Software* (2023)).

Traditional translation models, grounded in rule-based systems or statistical machine translation, have long been the bedrock of linguistic translation. Rule-based systems are constrained by predefined rules, while statistical models require parallel corpora for training. These approaches, while effective, face challenges in handling informal language and diverse linguistic nuances. ChatGPT's approach to translation, rooted in unsupervised learning on a grand scale, represents a departure from traditional methodologies. Its adaptability to diverse language pairs and informal language use sets it apart. While not explicitly designed for translation, researchers have explored fine-tuning ChatGPT for specific translation tasks, showcasing its potential in challenging the status quo.

Beyond its out-of-the-box capabilities, ChatGPT's fine-tuning potential allows researchers to tailor the model for specific tasks, including translation. The methodology of fine-tuning involves exposing the model to task-specific data, refining its performance and responsiveness to domain-specific nuances.

The journey of ChatGPT as a translation model is an ongoing narrative, with researchers and developers actively exploring avenues to refine and enhance its translation capabilities. Ongoing efforts focus on improving accuracy, handling rare languages, and addressing specific domain challenges. As these trajectories unfold, ChatGPT and similar models are poised to further influence the landscape of translation technologies.

In conclusion, ChatGPT's evolution from a conversational agent to a dynamic translation model represents a transformative chapter in the narrative of NLP. Its adaptability, contextual understanding, and defiance of traditional translation norms mark a

paradigm shift. As researchers continue to unravel the intricacies of ChatGPT's capabilities, it stands poised to redefine the landscape of translation technologies, offering a glimpse into the future of AI-driven language understanding and generation.

This exploration sets the stage for the subsequent sections of this thesis, where we will delve deeper into the methodology of fine-tuning ChatGPT and NLLB200 for language translation tasks. Through meticulous experiments, we aim to uncover the nuances and potential of ChatGPT and NLLB200 as translation models.

Chapter 3

Methodology

3.1 Hypothesis

Before commencing the experiments, we develop hypotheses based on existing knowledge, observations, or theoretical frameworks. These hypotheses pose specific statements or predictions about the relationship between variables. Our hypothesis was formed after careful background research, and formulated in section 1.4

3.2 Experimental design

We will construct five UG-defying rules, with different levels of disruption of statistical regularities in the surface form, and apply these rules to twenty Nynorsk sentences. The sentences will be partly constructed manually, and partly constructed with the help of artificial intelligence.

The experiments in this thesis will be split into two categories: Human participants and machine learning models. In the human participants' experiment, ten individuals will be asked to translate two sentences from five different universal grammar-defying rules, for a total of ten translations. The participants will not have any prior knowledge of the UG-defying rules applied to the sentences provided.

The machine learning models will be asked to translate one hundred sentences each, with different prior knowledge of the rules. Some models will be tested on 0-shot experiments, while others will be tested on a 1-shot experiment.

3.3 Variables and controls

In the five UG-defying rules we will create, there will be two rules which will disrupt statistical regularities in the surface form on a basic level, two which will do so moderately and one rule that will invert the linear order of words in a sentence completely.

We will construct one hundred grammatically legal Nynorsk sentences, twenty for each rule, where ten will have a normal conventional significance, and ten absurd sentences. This is done to prevent the machine learning models and human participants from guessing based on context. All one hundred sentences will be translated to English manually to serve as a target translation. Finally we will alter these sentences with

the five UG-defying grammatical rules to create the UG-defying sentences to be translated. The original Nynorsk sentences will be saved to act as a control for the machine learning models.

3.4 Sampling and participants

We decided ten participants for the human translation experiment would be the best option, as they would generate a total of one hundred translations which we could then compare to the machine learning models one hundred translations. We will create a simple questionnaire to send out to friends and family, which again will be encouraged to keep sharing until we will receive ten complete questionnaire results. This was done with the aim to create a random sample, ensuring that the results can be applied to a broader population.

The models we chose for our machine learning model experiments were based on cutting edge performance and overblown publicity of the models.

3.5 Data collection

The data we will use in this research are based on the survey the human participants answered and experiments with the machine learning models. The data we receive will be stored in dedicated spreadsheets for further analysis.

3.6 Data analysis

After gathering sufficient data for all experiments, we will analyze metadata such as mean and median scores, as well as creating bar charts and boxplots for ease of understanding and readability. We will also do a one-way ANOVA pairwise comparison to get a better understanding on whether or not the data is significant or not.

3.7 Ethical considerations

In conducting this master's thesis, ethical considerations play a paramount role in ensuring the integrity, respect, and protection of participants' rights. The study will adhere to the principles of informed consent, whereby participants will receive comprehensive information about the research objectives and procedures before agreeing to participate. Confidentiality measures will be rigorously implemented to safeguard the privacy of participants, and all data will be anonymized to prevent the identification of individuals.

Chapter 4

Methods

4.1 Creating universal grammar-defying sentences

The experimental process commenced with the creation of rules defying universal grammar, influenced by the Nynorsk lexicon as a familiar foundation. This selection was driven by our comfort with Nynorsk, which facilitated the generation of sentences and translations without external assistance. A significant rationale behind this choice was to explore the use of a low resource language as a basis for our sentences that break the principles of universal grammar. This decision was informed by the hypothesis that machine learning models excel in low resource languages due to their acquisition of universal grammar principles from other high resource languages. To maintain the novelty of our fake language, it was imperative that the rules we devised had no existing analogues in other languages, although we acknowledge our non-linguist status and the potential for undiscovered linguistic parallels.

The creation of our universal grammar-defying sentences involved formulating five rules designed to deviate from universal grammar principles. The first rule, termed "sentence final negation," mandated that the word "ikkje" always appear either after the last word before a comma or at the end of the sentence. For instance, the sentence "Eg vil ikkje eta grønsaker" would become "Eg vil eta grønsaker ikkje" under this rule.

The second rule, referred to as "inverse linear order," required reversing the linear word order in a sentence. This resulted in the fake language sentence reading in reverse compared to the original sentence e.g I[1] Noreg[2] kan[3] du[4] oppleve[5] midnattssol[6] om[7] sommaren[8] becomes Sommaren[8] om[7] midnattssol[6] oppleve[5] du[4] kan[3] Noreg[2] i[1].

The third rule, named "passive conjugation," involved conjugating every verb in its passive form, departing from correct grammar. An implementation of this rule on the sentence Han køyrer ein svart bil would result in Han køyrast ein svart bil.

The fourth rule, denoted as "noun precedes adjective," transposed the positions of nouns and adjectives. In Nynorsk, adjectives typically precede nouns, but this rule stipulated the reverse arrangement: Den harde[adj.] isen[noun] knakar under støvlane yields the sentence Den isen[noun] harde[adj.] knakar under støvlane.

The fifth and final rule, "verb direct object swap," entailed interchanging the positions of verbs and direct objects in a sentence. When the rule is applied to the sentence Han kjøper[verb] ei jakke[dir. obj.], the result will be Han ei jakke[dir. obj.] kjøper[verb].

Upon formulating these rules, the subsequent step encompassed determining the number of sentences to generate. A total of twenty sentences were selected for each rule, yielding a comprehensive dataset of one hundred sentences. Within each rule category, ten sentences adhered to conventional grammatical patterns, while ten veered into unconventional and eccentric territory. This diversity aimed to prevent machine learning models from deducing context based solely on word recognition.

To ensure uniformity and correctness, the sentences underwent manual refinement, with particular attention paid to maintaining adherence to the newly established rules. The original grammatically correct Nynorsk sentences were then systematically translated into English to serve as target translations for performance evaluation.

The organized compilation of these sentences, both in fake language and their corresponding correct English translations, was preserved within a structured spreadsheet, categorized by rule for ease of navigation and comparison. This spreadsheet layout facilitated streamlined evaluation of machine learning model translations against their intended targets.

4.2 NLLB200

The initial steps involved importing the necessary components, AutoModelForSeq2SeqLM and AutoTokenizer, from the transformers library. Subsequently, the model, specifically the "nllb-200-distilled-600M" variant, was loaded from the huggingface repository. Although larger NLLB models exist, practical limitations led to the selection of this particular variant that could be accommodated within the available computational resources. Upon loading the model, the source and target languages, Nynorsk and English in this context, were specified. Subsequently, the model was provided with a sentence, referred to as an "article," and it produced a corresponding translation.

Following preliminary trials involving fabricated Japanese sentences from Musso et al.'s work on Broca's area and language instinct, our focus shifted to the constructed fake Nynorsk sentences. Given varying outcomes in the initial Japanese sentence tests, a meticulous quality control process was established to ensure the absence of anomalous outputs. This involved subjecting the model to individual sentences in a meticulous manner, albeit involving the repetitive task of testing a hundred sentences. Despite its labor-intensive nature, this approach served to validate the setup process and confirm the model's predictable behavior.

The subsequent step involved a targeted examination of the model's response to grammatically erroneous sentences, beginning with a set of twenty sentences adhering to the sentence final negation rule. This subset allowed us to gain insights into the model's reaction to syntactically invalid constructions. Once the model's behavior was verified, the evaluation continued with the remaining sentences. The resulting translations generated by the NLLB200 model were systematically recorded within a spreadsheet alongside legally constructed Nynorsk sentences, the sentences warped by our universal grammar-defying rules, and the intended target translations. This organization facilitated convenient navigation and streamlined the process of assessing translation performance. Moreover, the side-by-side arrangement of NLLB200 translations and manually translated target sentences offered a rapid overview of the model's performance on specific sentences.

4.3 ChatGPT

During the experimentation phase involving NLLB200, there was a notable surge of attention directed towards ChatGPT across conventional and digital media platforms. Encouraged by our experience with NLLB200's translation capabilities, we embarked on a comparative study involving ChatGPT to assess its performance against Meta AI's cutting-edge translation model.

It's important to acknowledge that while ChatGPT has the potential for language translation, it is not primarily designed for this purpose, unlike NLLB200. Nonetheless, ChatGPT possesses intricate pattern learning capabilities, prompting our curiosity regarding its proficiency in comparison to a state-of-the-art translation model.

To ensure equitable evaluation between the models, we devised a strategy similar to that employed for NLLB200, albeit with a slight variation. Instead of presenting ChatGPT with sentences individually, we opted to provide five sentences simultaneously, each derived from different fake grammatical rules. This approach mitigated potential pattern recognition advantages and enhanced efficiency in data submission.

Creating a suitable prompt for ChatGPT necessitated precision in phrasing. Initially, we requested the model to translate the Nynorsk sentences into English without explicitly mentioning the manipulation by fake grammatical rules, mirroring the approach taken with NLLB200. However, this elicited a response indicating that the sentences were linguistically invalid, necessitating refined phrasing. After iterative experimentation, a more effective prompt was established: "Translate these Nynorsk sentences for me. Ignore the fact that they are not grammatically legal in Norwegian, but you have to translate them to a grammatically legal English sentence." This prompt, along with the distinct set of five sentences, formed the input for ChatGPT's translation task. The resulting translations were meticulously organized in a dedicated spreadsheet, mirroring the approach adopted for NLLB200.

Further exploration involved examining whether ChatGPT's performance could be enhanced through exposure to a single universal grammar-defying sentence and its target translation, along with simultaneous exposure to all twenty sentences from each grammatical rule. To ensure these cues were not part of the sentences to be translated, a new sentence and target translation were created for each rule. In the prompt designed for this variant of the experiment, we introduced the concept of a fabricated language based on Norwegian Nynorsk, provided an illustrative example, and instructed the model to disregard Norwegian grammatical conventions while generating a grammatically correct English sentence. Following this context, the model received the twenty sentences corresponding to the designated rule.

The outcomes of this one-shot experiment with ChatGPT were methodically cataloged in a dedicated spreadsheet, mirroring the organization methodology employed for the prior experiments involving NLLB200 and ChatGPT.

4.4 Human Translation

A crucial aspect of our study aimed to contrast the performance of human participants with the generated AI models on the sentences deviating from universal grammar. To accomplish this, we devised a concise questionnaire consisting of ten sentences, en-

compassing two sentences for each of the five rules. Sentence selection for the questionnaire was influenced by the scoring of NLLB translations, guiding our choices based on these scores. In certain instances, we opted for sentences with similar scores but that carried greater intrinsic interest. Ideally, we aimed for human participants to translate all one hundred sentences; however, considering potential participant fatigue, we settled for ten sentences, hoping to gather a minimum of ten participants for comparison against the AI models.

The questionnaire commenced with an introduction outlining the experiment's premise. Participants were informed that they would encounter two sentences for each of the five rules from a larger pool of one hundred sentences. Clear instructions emphasized participants' liberty to reference Nynorsk and English vocabulary as needed and assured the presence of grammatically correct target translations. The participants remained unaware of the specific fake grammatical rules and the sequence in which sentences were presented.

The sentence selection process began with the "sentence final negation" rule. Participants were presented with sentences "Ho vil at eg skal seie det til nokon ikkje" and "Sjølv om eg likar å ete bananar ikkje, har eg alltid ein med meg." This pair effectively exemplified the rule's operation, featuring a basic application in the first sentence and a more intricate instance involving comma placement in the second.

For the "inverse linear order" rule, participants encountered sentences "Sommaren om midnattssol oppleve du kan Noreg i" and "Elefant rosa ein med poker spele å likar min katten." These sentences, although potentially perplexing at first glance, were designed to be decipherable through human intuition, capitalizing on pattern recognition.

The "passive conjugation" rule was represented by sentences "Geitene spelast sjakk på bordet i hagen" and "Dei byggast ein katedral av pølser og spaghetti." Both sentences exhibited verbs conjugated in their passive form ("spelast" and "byggast"), deviating from conventional grammar to challenge participants.

Sentences "Den fjellveggen lange er dekorert med prikkar store rosa" and "Den sola lilla er så stor at ho dekkjer heile himmelen" encapsulated the "noun precedes adjective" rule. The former sentence introduced dual occurrences of the rule, requiring translation adjustments for both "Den lange fjellveggen" and "store rosa prikkar." The latter sentence illustrated a single instance of the rule, transforming "Den lilla sola" to "Den sola lilla."

The "verb direct object swap" rule was represented by sentences "Eg såg ein frosk som ein ATV køyrde gjennom skogen" and "Kaninen min piano spelte for ein gjeng med sjiraffar." Although straightforward, these sentences incorporated illogical phrasing.

Upon selecting the ten sentences, we distributed the questionnaire to acquaintances, encouraging them to share it within their social circles. Our goal was to secure a minimum of ten participants, cognizant of the potential for differing interpretations of the task's objectives. The collected responses were intended to provide a meaningful human performance benchmark, albeit with inherent limitations in terms of the sentence variety compared to the AI models' dataset.

Chapter 5

Results and Discussion

5.1 Early results and adjustments

An analysis of the results from our experiment with NLLB200 showed a varying score across the different rules. NLLB200 was given one sentence which violated the principles of universal grammar at a time to translate, and the translation was then stored in a spreadsheet (see methods). The model was only instructed to translate from Nynorsk to English, with no knowledge that the sentences had been altered by our universal grammar-defying rules. Upon running the scoring program on the one hundred sentences the algorithm had translated, we encountered our first issue with the scores. We had done some cleaning of the data in the form of lowercasing the data and removing punctuations, but upon looking closer at the scores we noticed that some of the translations were very harshly scored compared to what we believed to be a fair score. One example of this is where NLLB200 translated the sentence Ho kjem til konserten ikkje" to She doesn't come to the concert with the target translation being Shes not coming to the concert. This translation only scored 0.20 with BLEU. In general the BLEU scores were lower than we expected across the board, with NLLB200 only having five out of twenty sentences scored higher than 0.1 on the sentences warped by the inverse linear order rule. While we expected subpar results in this bracket of sentences, the scores seemed to not give a good representation of the translations.

Following the evaluation of the scores provided with BLEU, we decided to look for a better solution for scoring, and after some research we found that the BLEU score is designed to be a corpus measure and it has some undesirable properties when used for single sentences. The Google BLEU score, or GLEU, was our solution to the problems we were having with the BLEU score, as the purpose of the Google BLEU score is to mitigate these undesirable properties when applied to individual sentences. In computing this score, all subsequences of 1, 2, 3, or 4 tokens in both the output and target sequences (n-grams) are logged. The subsequent calculation involves computing precision and recall, as defined:

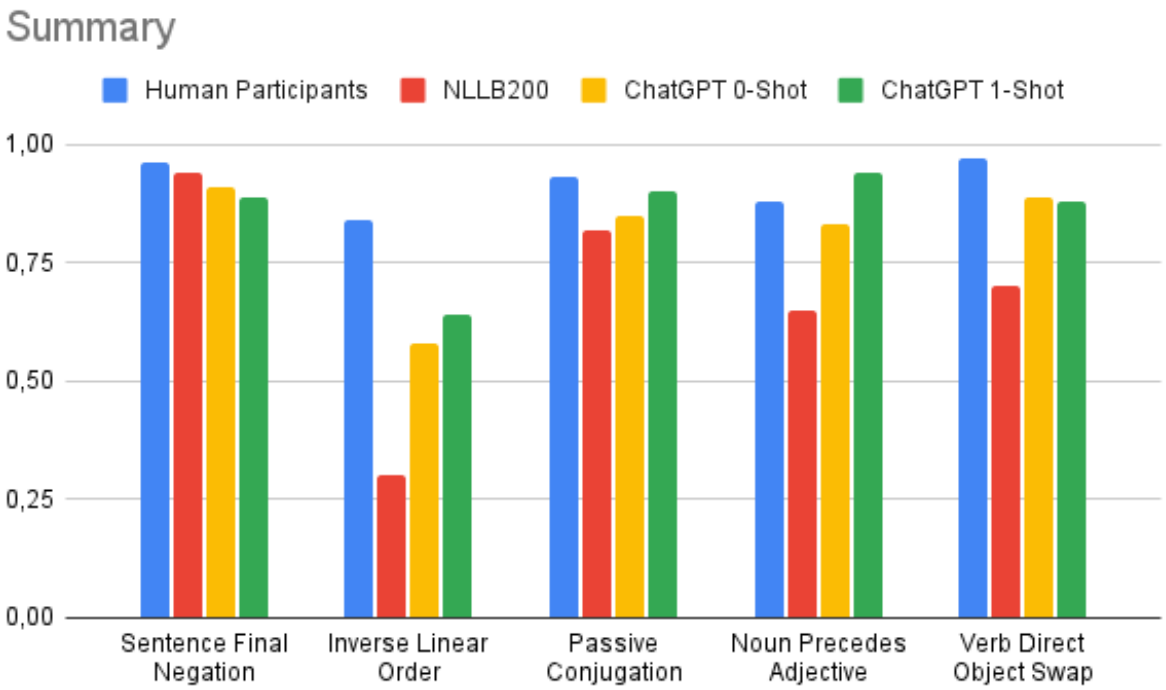
- Precision: the ratio of matching n-grams to the total number of n-grams in the generated output sequence.

- Recall: the ratio of matching n-grams to the total number of n-grams in the target (ground truth) sequence.

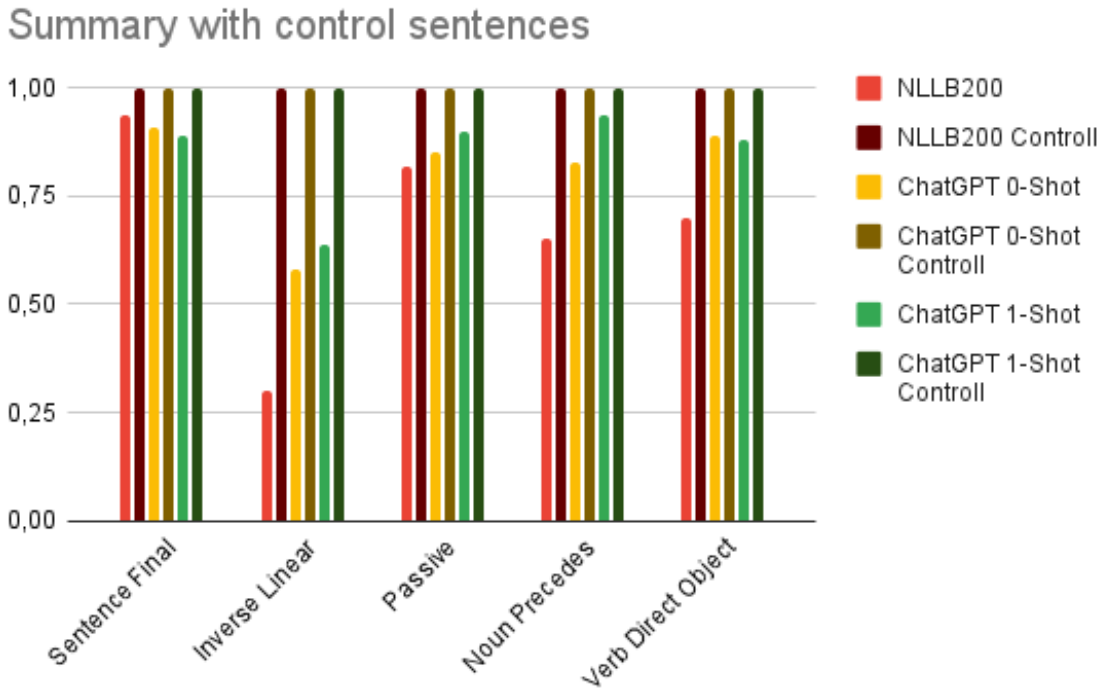
The score is determined by returning the minimum value between precision and recall. Typically employed for assessing machine translation models, this metric proves especially valuable when examining scores for individual (prediction, reference) sentence pairs, rather than averaging over the (prediction, reference) scores for an entire corpus. However, it can also be applied in scenarios where averaging across the scores for an entire corpus is required (*Wu et al.* (2016)).

Another issue we encountered early in the process was the choice of words by the translation models. As there was only one correct translation for each sentence, the phrasing by the models would have to be very precise and use the exact same vocabulary that was employed when creating the target translation. The English language is, however, not so simple, and simple differences in words from American English and British English could result in a lower score for the models. To address this we methodically reviewed the translations to look for near-synonymous words or sentences and adjusted the translations before running them through the scoring process. Some examples of near-synonymous or synonyms include but are not limited to spaceship and spacecraft, hut and cabin, lunar and moon (eclipse), stinking and smelly, band and bunch (of giraffes). There were also more complex cases where there was not only one single synonymous word that was different. An instance of this was They are building a cathedral made out of sausages and spaghetti with the target translation being They're building a cathedral out of sausages and spaghetti where the model is translating it to a more polite phrasing with They are and made out of instead of Theyre and out of. The initial translation only received a GLEU score of 0.32, which would be very harsh considering that the translation has the same meaning that the target translation has. We therefore decided to adjust these sentences that meant the same to the target translation, meaning that this example sentence which originally received a score of 0.32, now receives a score of 1.0. This method for adjusting near-synonymous words and sentences was applied to all translations from both human participants and machine learning models.

5.2 Final results



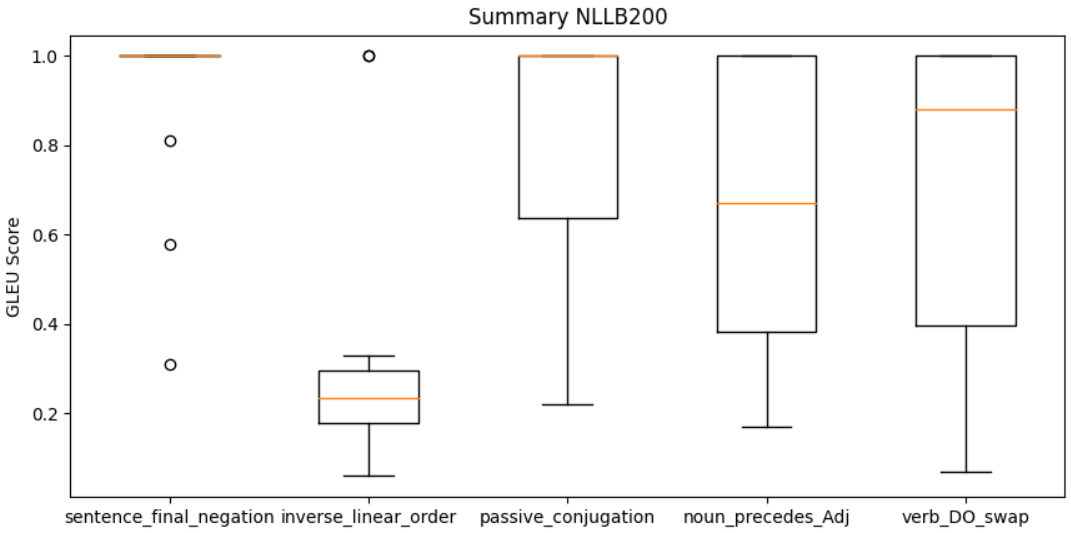
The figure above represents a bar chart that displays the mean score of every experiment for each rule, where blue represents the human participants, red represents NLLB200, orange represents ChatGPT 0-Shot and green represents ChatGPT 1-shot.



To check how the machine learning models would perform on the source language the universal grammar-defying sentences were based on, Nynorsk, we decided to conduct a small experiment where we would give the models all one hundred sentences,

before they were altered by our rules, as a control measure. The nature of the experiments would be identical to the universal grammar-defying sentence experiments, meaning NLLB200 would be given one sentence at a time, ChatGPT 0-shot five sentences at the same time but from different brackets and ChatGPT 1-shot all twenty sentences from the same bracket after given an example of a nynorsk sentence and the english translation. The translations would then undergo the same adjustments that were done in the main experiment. While the models in the main experiments would translate the sentence Bestemora mi lagar pizza ikkje, men laksesuppe med sjokoladebitar, they instead receive the sentence Bestemora mi lagar ikkje pizza, men laksesuppe med sjokoladebitar for this control experiment. Due to Nynorsk being a low resource language, we were uncertain about the performance quality of the models, but all three models passed with a perfect score for all one hundred sentences.

5.3 NLLB200



The boxplots that are included for each of the experiments shows a box which extends from the first quartile (Q1) to the third quartile (Q3) of the data, with the orange line representing the median. The whiskers reach from the box to the outermost data point contained within 1.5x the interquartile range (Q1 to Q3) from the box. Every data point that extends beyond the whiskers are outliers.

Following the changes made in the scoring process, we can see from the overview that NLLB200 handles the universal grammar-defying sentences warped by the sentence final negation rule very well with a mean score of 0.94, while struggling with the inverse linear order sentences with a mean score of 0.3. It is also worth noting that while the results are not horrible, the scores from the noun precedes adjective sentences are subpar, featuring a mean score of 0.65, closely followed by verb direct object showcasing a mean score of 0.7. NLLB200 also achieved a respectable mean score of 0.82 in the passive conjugation bracket. The boxplot shows us that there was greater variability in the GLEU scores for passive conjugation, noun precedes adjective and verb direct object swap than inverse linear order and sentence final negation, while the two latter had greater outliers.

In the sentence final negation segment of the sentences, NLLB200 achieved a flawless 1.0 score on all sentences except three, and was where this model performed the best with a solid 0.94 average GLEU score. Boxplot name also confirms this with a very low distribution of data points, with the three imperfect translations being the outliers. The worst rated translation this model did in this rule bracket derive from the sentence Me må gløyme å betale rekningane ikkje which NLLB200 translated to We have to forget to pay the bills, while the correct translation would be We must not forget to pay the bills, completely ignoring the negation in the sentence resulting in a GLEU score of 0.58. The second error the machine learning model made was when translating Ho vil at eg skal seie det til nokon ikkje to She wants me to tell someone who doesn't. This time the negation is included in the result, but the translation ended up lacking order, resulting in a GLEU score of 0.31 with the target translation being She doesn't want me to tell anyone. The final translation to receive a less-than-perfect score was a result of the sentence Treet i hagen min har blad ikkje, men rosa sukkerspinn which NLLB200 translated to The tree in my garden has no leaves, but pink sugar spines which was close to the target translation, The tree in my garden has no leaves, but pink cotton candy, yielding a 0.81 GLEU score.

The results from the inverse linear order section were pretty bad, with only two translations receiving a GLEU score greater than 0.35. The worst rated translation for this rule came from the sentence Fjorder vakre sine for kjent er Noreg and was translated to The forefathers of their beauty are too well known: Norway with the correct translation being Norway is known for its beautiful fjords. This translation only received a GLEU score of 0.06. The second worst translation was from the sentence Luft frisk pustar og fjellet i går eg which was translated to The air breaks and the mountain yesterday with the target translation I walk in the mountains and breathe fresh air, yielding a GLEU score of 0.1. The third worst translation received a GLEU score of 0.13, with the original sentence being Landet heile over kjent er pinnekjøtt og rakfisk som mat Norsk translated to The whole country is known for its pollen and raccoon with the goal of translating to Norwegian food such as rakfisk and pinnekjøtt is known all over the country. While most of the scores in this bracket was between 0.2 and 0.3, there were also two perfect 1.0 scores in this bracket. When asked to translate Noreg i hovudstaden er Oslo and Sommaren om midnattssol oppleve du kan Noreg i NLLB200 correctly translated these sentences to Oslo is the capital of Norway and In Norway you can experience the midnight sun in the summer.

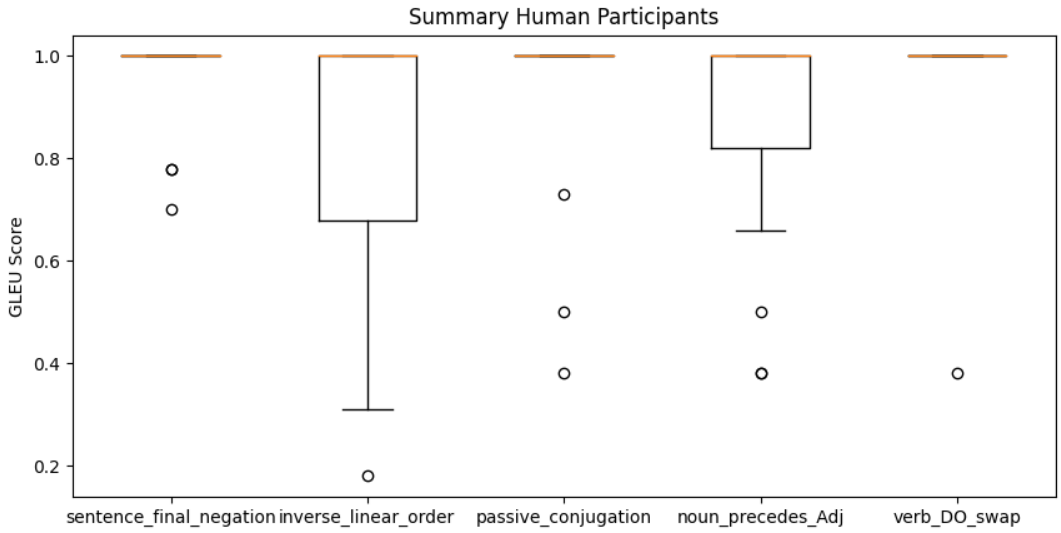
The passive conjugation rule was NLLB200s second best category in this experiment with ten perfect scores out of twenty. Most of the mistakes in translation in this bracket was due to NLLB200 translating the words in past tense. The sentence Dei spelast fotball på stadion being translated to They played football at the stadium and the correct translation being They're playing football in the stadium only received a GLEU score of 0.22, being the worst rated translation in this bracket. The translation with the second lowest quality, and the only other translation to score below 0.5 in this set of rules, stem from the sentence Ho sjåast ein vakker solnedgang translated to She looks like a beautiful sunset while the target translation was She sees a beautiful sunset, resulting in a 0.39 GLEU score. The third least satisfactory translation find its source from the sentence Sauene ridast på syklar nedover bakken, and NLLB200 translating this to The sheep ride bicycles down to the ground with the target translation being The sheep ride bicycles down the hill yielding a 0.58 GLEU score. A couple of examples

of correctly translated sentences scoring 1.0 were Han springast rundt i gatane med ei melon på hovudet which was translated toHe's running around the streets with a melon on his head. and Eg hoppast over månen med ein stige translated to I jump over the moon with a ladder.

In the noun precedes adjective section of the sentences, NLLB200 achieved a mean score of 0.65, being the worst rated rule in this experiment besides inverse linear order. The three worst rated translations out of the twenty in this bracket received the scores 0.17, 0.18 and 0.26 while also getting seven perfect scores of 1.0. The worst rated translation came from the sentence Den isen glatte gjer det vanskeleg å helde seg på beina translated to The ice is smooth, making it hard to stand on your legs" and the correct translation being The slippery ice makes it difficult to stay on your feet" which as previously stated received a GLEU score of only 0.17. The second worst translation with a score of 0.18 originated from Den fjellveggen lange er dekorert med prikkar store rosa" translated to The mountain walls are long, with dots of pink and the target being The long mountain wall is decorated with large pink dots. The third poorest translation came from the sentence Det treet høge kastar ei skugge lang over vegen which NLLB200 translated to The tree tall throws a shadow long across the road with the goal being The tall tree casts a long shadow over the road. Den sola gule står opp over horisonten" and Den elefanten rosa går elegant på tå are examples of translations the model got right with the translations The yellow sun rises above the horizon and The pink elephant walks elegantly on tiptoe.

NLLB200s attempt to translate sentences warped by the verb direct object rule was for the most part very good, with ten translations securing a 1.0 GLEU score, but the mean score suffered from one of the translations only scoring 0.07 and two translations receiving a GLEU score of 0.26. The 0.07 score originated from the sentence Han mora si helsa på" which was translated to He's got his mother's health and the correct translation being He greeted his mother, which was probably caused by confusion around the word helsa which can mean both health and greeted when used in different contexts. Ein robot ein biff åt med ei gaffel og kniv was the source for the first 0.26-score translation, where NLLB200 translated the sentence to A robot, a steak, a gaffel and a knife while the target translation was A robot ate a steak with a fork and knife. The second translation to score 0.26 was Kua mi fransk snakka medan ho eit fiolinstykke spelte" which was translated to My mother was speaking French while she was playing a violin while My cow spoke French while playing a violin piece was the correct translation. Some correct translation where Krokodillen min ein bok las om romfart på stranda and Ein skilpadde eit maraton sprang på under ti minutt which both received a perfect GLEU score of 1.0 with My crocodile read a book about space travel on the beach and A turtle ran a marathon in less than ten minutes".

5.4 Human participants



We analyzed data from ten participants that had participated in the questionnaire. They were unaware of the nature of the different rules before trying to translate the universal grammar-defying sentences (see Methods). Because the ten human participants only translated ten unique sentences (two from each rule), compared to the machine learning models who translated one hundred unique sentences, a direct comparison would be challenging. But as the participants would translate a total of one hundred sentences, the results they produced would give us a good baseline to judge the models on. The human participants received a very good score across all five rules, with verb direct object swap and sentence final negation being the highest scoring brackets of translations with a mean score of 0,97 and 0.96, while inverse linear order was the worst with a respectable mean score of 0.84. Noun proceeds adjective was slightly better with a mean score of 0.88, while passive conjugation received a mean score of 0,93 with the human participants. As shown by Boxplot name, most of the scores in this experiment resulted in a perfect 1.0, and the median on all categories of rules being 1.0. There is also very little variation in the data points, other than in the inverse linear order and noun precedes adjective bracket.

In the first rule the human participants encountered, sentence final negation, seventeen translations scored a perfect 1.0, while the worst translation received a GLEU score of 0.7. This translation came from the sentence Sjølv om eg likar å ete bananar ikkje, har eg alltid ein med meg where the participant translated the sentence to Even though I like eating bananas, I don't always have one with me aiming to translate into Even though I don't like eating bananas, I always have one with me, missing out on a perfect score as the participant placed the negation in the wrong part of the sentence. The two other imperfect translations originate from the same sentence, and both of the participants made the same error when trying to translate it. Both participants neglected the eating part of the sentence, resulting in a translation that indicates the person dont like bananas in general, instead of the dislike for eating them. The translation the participants submitted was Even though I dont like bananas, I always have one with me, resulting in two translations scoring 0.78 on the GLEU score.

The inverse linear order sentences received a wider range of scores for the translations, but the participants still managed to get thirteen out of the twenty sentences correct. The worst translation in this bracket was from the first sentence the participants were introduced to for this rule, Sommaren om midnattssol oppleve du kan Noreg i, where one of the participants submitted the translation The summer midnight sun is exclusive to Norway" and the target being In Norway you can experience the midnight sun in the summer resulting in a GLEU score of only 0.18. The next worst translation, scoring 0.31, derived from the other sentence the participants were asked to translate for this rule, Elefant rosa ein med poker spele å likar min katten, which in was translated to A pink elephant, one with a poker game, also likes my cat with the target translation being My cat likes to play poker with a pink elephant. Four other participants struggled with this sentence, as all of them translated it to A/The pink elephant likes to play poker with my cat resulting in one score of 0.59 and three 0.68 scored translations. The final translation that did not receive a perfect score came from the midnight sun sentence, where one of the participants translated the sentence to In Norway you can experience the midnight sun, missing the in the summer part of the sentence and receiving a 0.68 GLEU score.

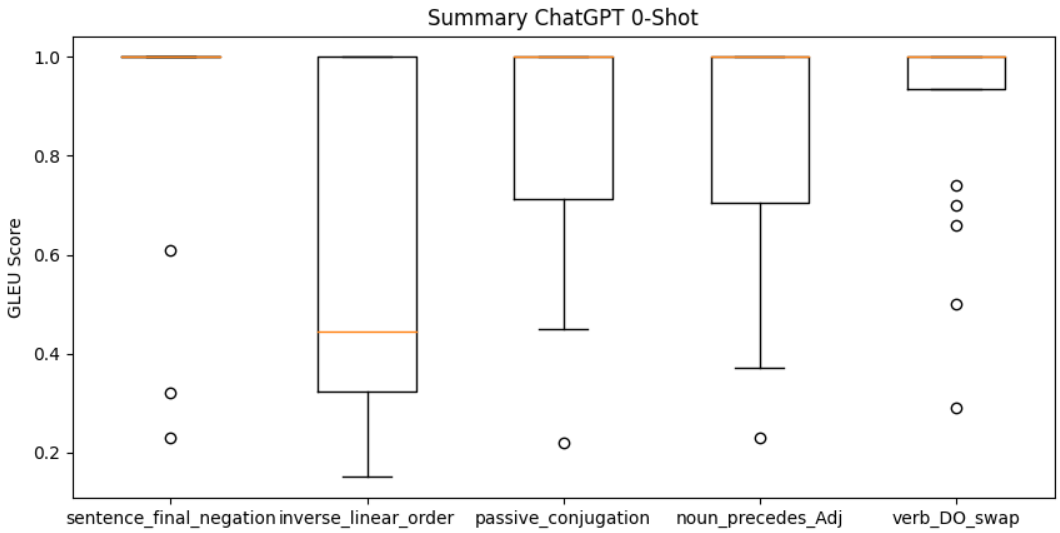
Passive conjugation was the third best bracket of sentences for the human participants, and they only received three scores which was not a 1.0. The first sentence the participants were asked to translate from this rule was Dei byggast ein katedral av pølser og spaghetti, and is also where all three imperfect scores originate from. The worst score comes from the translation A cathedral made of sausages and spaghetti is being constructed while the target translation was They're building a cathedral out of sausages and spaghetti " resulting in a GLEU score of 0.38. The second worst translation is only slightly better with a 0.5 GLEU score: A cathedral is being built out of sausages and spaghetti. The translation They built a cathedral out of sausages and spaghetti received a score of 0.73. The second sentence the participants were asked to translate for this rule was Geitene spelast sjakk på bordet i hagen with everyone achieving a perfect score by submitting the translation The goats are playing chess on the table in the garden.

The second worst category for the human participants was noun precede adjective, but they still received fourteen perfect scores for their translations. The first sentence the participants were asked to translate in this category was Den fjellveggen lange er dekorert med prikkar store rosa with the target translation being The long mountain wall is decorated with large pink dots. Four of the six imperfect scores came from this sentence, with the worst scoring translation being That long rock wall is decorated by large pink dots receiving a GLEU score of 0.38, while the second worst, The tall mountain wall is decorated large pink dots received a score of 0.5. The other two translations that were not perfect from this sentence was The tall mountain wall is decorated with large pink pink dots receiving a score of 0.66 and The mountain wall is decorated with large pink dots with a GLEU score of 0.79. The second sentence the participants were introduced to for this rule was Den sola lilla er så stor at ho dekkjer heile himmelen with the target translation The purple sun is so big that it covers the whole sky. The participants did well with this sentence, but the translation The purple sun is large enough to cover the whole sky only received a 0.38, while another participant presented the translation The pink sun is so big that it covers the whole sky missing out on a perfect score with 0.83 because of using the word pink instead of

purple to describe the sun.

Verb direct object swap was where the human participants had the best mean score, with only one translation not being scored as a perfect 1.0. The subpar translation derive from the first sentence the participants were introduced to in this rule bracket:Eg såg ein frosk som ein ATV køyrde gjennom skogen with the desired translation being I saw a frog driving an ATV through the forest. The sub par translation, and the only translation which scored less than 1.0 in this section, was I saw an ATV which a frog was driving through the forest receiving a score of 0.38. The subsequent sentence the human participants were met with was Kaninen min piano spelte for ein gjeng med sjiraffar where the correct translation was My rabbit played piano for a bunch of giraffes. While there originally were a couple of non-perfect scores from this sentence, they were eventually scored as a 1.0 after going through the adjustment process mentioned earlier. A lot of misspelling of the word giraffe and a couple of users using the word group instead of bunch were the main problems with the translations for this sentence.

5.5 ChatGPT 0-shot



Following the experiments done with Meta AIs NLLB200 model and getting a good benchmark with the human participants, OpenAIs ChatGPT was next. The nature of the experiment was comparable to the methods used with NLLB200, as the model was presented with one sentence from each rule at the same time, which in turn mitigated potential pattern recognition advantages and enhanced efficiency in data submission. ChatGPT was presented with the exact same universal grammar-defying sentences that NLLB200 was presented with. In this 0-shot experiment, ChatGPT performed better than expected in most of the categories, with a mean score of 0.91 in sentence final negation, 0.89 in verb direct object swap, 0.85 in passive conjugation and 0.83 in noun precedes adjective. ChatGPT also performed adequately on the inverse linear order sentences with a mean score of 0.58. From Boxplot name we can see that the median score from this experiment is 1.0 on all categories except inverse linear order. Sentence final negation and verb direct object swap have little variation in the scores, with three and five outliers each, while the other three categories have a greater variability in

GLEU score.

In the ChatGPT 0-shot experiment, the machine learning model did very well by scoring a perfect 1.0 with all translations other than three, all being considered as outliers. The worst rated translation ChatGPT made in this section of the rules originate from the sentence Katten min vil ete fisk ikkje, berre bananar with the target translation being My cat won't eat fish, only bananas, and ChatGPT translating to My cat wants to eat fish not just bananas receiving a GLEU score of only 0.23. The second least precise translation traced back to the sentence Ho vil at eg skal seie det til nokon ikkje, where ChatGPT translated it to She wants me to tell someone not while the correct translation would be She doesn't want me to tell anyone. For this translation ChatGPT simply placed the negation where it was in the source sentence, resulting in a 0.32 GLEU score. ChatGPT repeated this mistake for the last imperfect score in the sentence final negation rule bracket, where the source sentence was Eg vil eta grønsaker ikkje. While I don't want to eat vegetables would be the correct way to translate this sentence, ChatGPT translated it to I want to eat vegetables not resulting in a 0.61 GLEU score. Some examples of correctly translated sentences in this bracket was Eg kan sjå på TV ikkje, for eg er redd for at den vil bite meg and Hunden min er ein hund ikkje, men ein alien frå Pluto which the model correctly translated to I cant watch TV because Im afraid it will bite me and My dog is not a dog but an alien from Pluto.

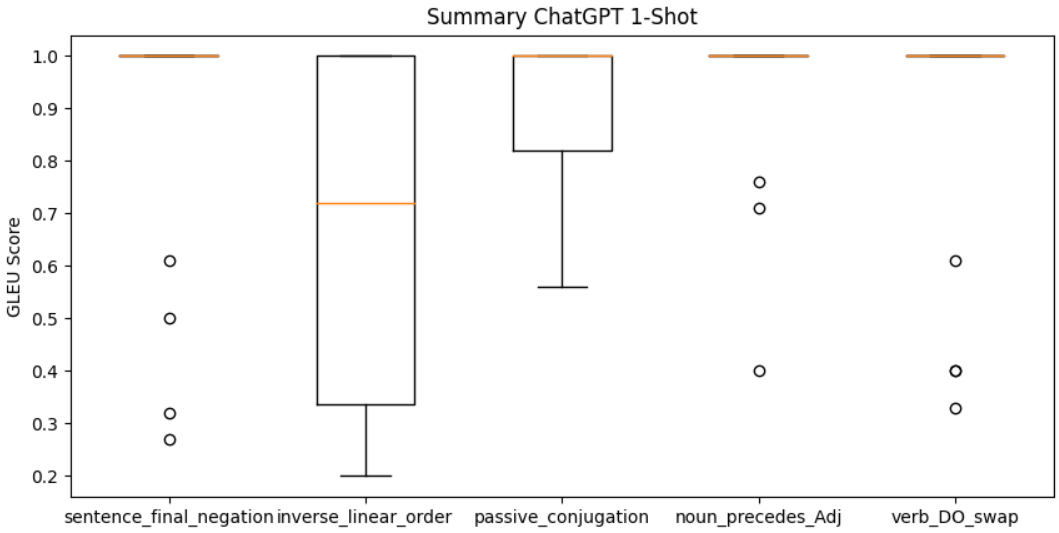
Inverse linear order was the worst category in the ChatGPT 0-shot experiment, but it was not a bad effort by ChatGPT with a 0.58 mean GLEU score. In this rule bracket, the model managed to receive six perfect scores, with the lowest score being 0.15. This translation has its origin from the sentence Meg med kvantemekanikk diskutere å likar og astrofysikk i doktorgrad ein har min sauen while ChatGPT translated it to I enjoy discussing quantum mechanics and astrophysics in my PhD and the correct translation being My sheep has a PhD in astrophysics and likes to discuss quantum mechanics with me. Two translations received the second lowest score of 0.26, with the first coming from the sentence Landet heile over kjent er pinnekjøtt og rakfisk som mat Norsk being translated to The whole country is known for Norwegian dishes such as pinnekjøtt and rakfisk with the target translation being Norwegian food such as rakfisk and pinnekjøtt is known all over the country. The second translation to receive a 0.26 was from the sentence Stylter på går han medan gitar spele å likar min sjiraffen being translated to While playing the guitar my giraffe walks on stilts and the correct translation being My giraffe likes to play the guitar while walking on stilts. A couple of examples the model got right in this rule was from the sentences Månen på pingvin ein med salsa alltid dansar eg and Elefant rosa ein med poker spele å likar min katten which was translated to I always dance salsa with a penguin on the moon and My cat likes to play poker with a pink elephant. In the passive conjugation section of the sentences to be translated, ChatGPT 0-shot managed to get a mean score of 0.85 with thirteen perfect translations. While most of the non perfect scores received a decent score, one translation stood out with a GLEU score of only 0.22. This translation originated from the sentence Dei spelast fotball på stadion while the translation was They play football at the stadium and the target translation being They're playing football in the stadium. The only other translations that received a GLEU score of less than 0.6 was one being scored as 0.45 and the other 0.56. The translation scoring 0.45 derive from the sentence Sauene ridast på syklar nedover bakken, and ChatGPT translated this to The sheep ride bicycles downhill with the target being The sheep ride bicycles down

the hill. The other translation came from the sentence Ho lagast god mat til middag which was translated to She prepares good food for dinner while the correct translation would be She's cooking good food for dinner. A few instances of a perfect translation include the sentences Dei byggast ein katedral av pølser og spaghetti and Ho snakkast med froskar på bunnen av dammen being correctly translated to They're building a cathedral out of sausages and spaghetti and She's talking to frogs at the bottom of the pond.

The next rule, noun precedes adjective, was the second worst rated category of translations in this experiment but with a respectable mean score of 0,83. ChatGPT managed to get eleven perfect translations, and only three of the translations were below 0.6 on the GLEU score with one of them being a 0.59. The least satisfactory translation stems from the sentence Den sykkelen rustne er ikkje til å stole på, which was translated to The rusty bike is not reliable with the target translation being The rusty bike is not to be trusted resulting in a GLEU score of 0.23. The second worst rated translation is associated with the sentence Den morgonenen lyse er full av forventning og håp which was translated to The morning is bright with anticipation and hope and the correct translation being The bright morning is full of anticipation and hope resulting in a 0.37 GLEU score. The final translation scoring less than 0.6 in this section of the experiment originate from the sentence Den katten vesle ligg og søv i vinduskarmen and was translated to The little cat is lying and sleeping on the windowsill by ChatGPT, while the correct translation would have been The little cat is sleeping on the windowsill. Samples of perfect translation include Den gulrota skrikande var ein sørgelig og ensom grønnsak translated to The screaming carrot was a sad and lonely vegetable and Den hytta blå er eit romskip forkledt som ein fjellhytte translated to The blue cabin is a spaceship disguised as a mountain cabin.

Verb direct object swap is the final rule in this experiment, and the model received a desirable mean GLEU score of 0.89. In this rule bracket, ChatGPT managed to achieve fifteen perfect translations, with only one translation scoring below 0.5 on the GLEU score. This translation originated from the sentence Han på konserten spelar gitar i kveled which was translated to He plays the guitar at the concert in the evening Having the desired translation as He's playing guitar at the concert tonight resulting in a 0.29 GLEU score. The second worst translation came from the sentence Eg såg ein frosk som ein ATV køyrde gjennom skogen, and was translated to I saw a frog riding through the forest like an ATV With the goal of translating to I saw a frog driving an ATV through the forest yielding a GLEU score of 0.5. The third least desirable translation from this bracket came from the sentence Ein hund swing dansa med ein sjimpanse i regnet which was translated to A dog danced the swing with a chimpanzee in the rain with the target translation being A dog was dancing swing with a chimpanzee in the rain achieving 0.66 on the GLEU score. Specific cases of perfect translations in this bracket include Elefanten min pingpong spelte med ein vassmelon and Kua mi fransk snakka medan ho eit fiolinstykke spelte being translated correctly to My elephant was playing ping pong with a watermelon and My cow spoke French while playing a violin piece.

5.6 ChatGPT 1-shot



The final experiment we did with our universal grammar-defying sentences involved the same machine learning model as in the previous experiment, ChatGPT, but this time we would first give the model an example sentence and the correct translation to the example sentence before asking it to translate all twenty sentences per rule at the same time in a 1-shot experiment. Overall the model performed quite well, even outperforming the human participants in the noun precedes adjective bracket with a mean score of 0.94. The model also performed well on passive conjugation with a mean score of 0.9, sentence final negation with a mean GLEU score of 0.89 and a mean score of 0.88 in the verb direct object swap bracket. This model also outperformed the other machine learning models in the inverse linear order, but was still worse than the human participants by a big margin with a mean score of 0.64. Looking at Boxplot name we can see that the GLEU scores from sentence final negation, noun precedes adjective and verb direct object swap were very concentrated at 1.0, with three or four outliers. The data points for passive conjugation were slightly more varied with no outliers, while inverse linear order is the only bracket with a median score lower than 1.0 with a 0.72 and a much greater variability in GLEU score compared to the others.

ChatGPT 1-shot surprisingly receives a lower mean score than ChatGPT 0-shot in the sentence final negation bracket with 0.89 compared to 0.91, this can however be caused by having one more outlier than in the previous experiment with four this time. The lowest-rated translation within this category is derived from the sentence "Katten min vil ete fisk ikkje, berre bananar", where the model translated to My cat wants to eat fish, not bananas with the correct translation being My cat won't eat fish, only bananas resulting in a 0.27 GLEU score. ChatGPT also did the same mistake in this 1-shot experiment that it did in the 0-shot experiment with the sentences Eg vil eta grønsaker ikkje and Ho vil at eg skal seie det til nokon ikkje with the translations I want to eat vegetables, not and She wants me to tell someone, not. On this occasion as well the model placed the negation at the end of the sentence, resulting in the same result as last time with the GLEU scores of 0.61 and 0.32. The new outlier in this experiment originate from the sentence Ho vil snakke med han ikkje, where the model again did the

same mistake by placing the negation at the end with the translation She wants to talk to him, not and the correct translation being She doesn't want to talk to him resulting in a GLEU score of 0.5. Representative cases of correct translations in this bracket include Bestemora mi lagar pizza ikkje, men laksesuppe med sjokoladebitar and Eg drikker kaffe ikkje, berre kaldt, grønt te-vatn being translated to My grandmother doesn't make pizza, but salmon soup with chocolate chips and I don't drink coffee, only cold, green tea water.

In the Inverse linear order bracket ChatGPT 1-shot outperformed ChatGPT 0-shot by 0.06 and NLLB200 by 0.34, but is still 0.20 points behind the human participants mean score. ChatGPT 1-shot achieved eight perfect translations in this bracket, but eight translation scoring 0.35 or lower reduces the mean score to 0.64. The two worst rated translations in this bracket derived from the sentences Meg med kvantemekanikk diskutere å likar og astrofysikk i doktorgrad ein har min sauen and Shakespeare lese å likar og russisk snakke kan mi kua, where the target translations were My sheep has a PhD in astrophysics and likes to discuss quantum mechanics with me and My cow can speak Russian and likes to read Shakespeare. ChatGPT 1-shot translated these sentences to I enjoy discussing quantum mechanics and astrophysics; I have a PhD in it. and I enjoy reading Shakespeare and speaking Russian; my cow can do it., resulting in a 0.2 score for both translations. The third least favorable translation came from the sentence Kristiansand i base har som fotballklubb ein er Start, which was translated to Kristiansand is the base of a football club called Start. with the target translation being Start is a football club based in Kristiansand yielding a 0.26 GLEU score. Illustrative examples of translations this model got a perfect GLEU score on were Mat handle skal eg når traktor rosa ein i rundt alltid køyrar eg and Dag dårleg ein har eg når mjølk i bade å elskar eg which the model correctly translated to I always drive around in a pink tractor when I go grocery shopping. and I love bathing in milk when I have a bad day.

ChatGPT 1-shot outperformed the other models with a mean GLEU score of 0.9 in the passive conjugation bracket, but were just short of the human participants 0.93 mean score. The model achieved twelve perfect translations and no translations scoring lower than 0.5, with the lowest rated score being 0.56. This translation originated from the sentence Dei spelast fotball på stadion, which was translated to They play football in the stadium. with the target translation being They're playing football in the stadium. The second worst translation came from the sentence Geitene spelast sjakk på bordet i hagen translated to Goats play chess on the table in the garden. with the correct translation being The goats are playing chess on the table in the garden resulting in a 0.61 GLEU score. The thirst least favorable translation originated from the sentence Eg flygast som ein fugl over fjelltoppane with the target translation being I'm flying like a bird over the mountain tops. ChatGPT 1-shot translated this to I fly like a bird over the mountain tops. resulting in a 0.73 GLEU score. A few instances of perfectly translated sentences are Katten mjauast høgt når ho er svolten and Eg trengast ein penn til å skrivast med which were translated to The cat meows loudly when she's hungry. and I need a pen to write with..

Noun precedes adjective was this model's best rule bracket with a mean GLEU score of 0.94, outperforming all the other models and the human participants. ChatGPT 1-shot achieved seventeen perfect translations, with only one of them being lower than 0.7. ChatGPT made the same error translating this sentence that it did in the 0-shot experiment, with Den katten vesle ligg og søv i vinduskarmen translated to The little

cat lies and sleeps on the windowsill. with the target being The little cat is sleeping on the windowsill. The use of lies and sleeps resulted in a 0.4 GLEU score. The second error the model made in this bracket was translating Den enga gule er full av einrar som syng i kor to The yellow meadow is full of birds singing in chorus. with the correct translation being The yellow meadow is full of reindeer singing in chorus resulting in a GLEU score of 0.71. The final imperfect translation the model made was using the word hammer instead of sledgehammer when translating Den puta harde kan brukast som ein slegge i nødstilfelle to The hard cushion can be used as a hammer in an emergency resulting in 0.76 on the GLEU score. Cases that demonstrate perfect translation are Den sykkelen stinkande luktar som ein blanding av søppel og bananar and Den elefanten rosa går elegant på tå translated to The smelly bike smells like a mixture of garbage and bananas. and The pink elephant walks elegantly on tiptoe..

In the final rule bracket, verb direct object swap, ChatGPT 1-shot performs very similarly to the 0-shot experiment, with a mean GLEU score of 0.88 compared to 0-shots 0.89. There are four outliers in this category, while the sixteen remaining translations are perfect. The worst rated translation received a 0.33 GLEU score with Han på konserten spelar gitar i kveled translated to He plays guitar at the concert in the evening. and the target being He's playing guitar at the concert tonight. There were also two translation receiving a 0.4 GLEU score in this bracket, both being translated with the verb conjugated in its past simple form instead of 3rd person singular form: Ho eit eple et and Han ei jakke kjøper was translated to She ate an apple. and He bought a jacket.. Instances that included perfect translations in this bracket were Vi ein deilig middag lagar til gjestane våre and Ein hund swing dansa med ein sjimpanse i regnet translated to We prepare a delicious dinner for our guests and A dog was dancing swing with a chimpanzee in the rain.

5.7 Pairwise comparison

To gain insight into the significance of these figures, we decided to do a one-way ANOVA pairwise comparison. For every pair of experiments compared, the one-way ANOVA will output one F(db) value, and a p value for every rule bracket compared. The difference between the rule brackets are significant if the p value reported is less than 0.05.

We started by comparing the human participant benchmark to NLLB200, with F(db) = 15.70. This comparison showed that there was no significant difference in the performance of NLLB200 and the human participants in the passive conjugation and sentence final negation rule brackets with a reported p of 0.8656 and 1.0 respectively. Noun precedes adjective was close to being significantly worse for NLLB200 than the human participants with p = 0.0708. NLLB200 did however perform significantly worse than the human participants in both inverse linear order and verb direct object swap brackets with a p value of 0.0 and 0.0128 each.

The F(db) value for the comparison between the human participants and ChatGPT 0-shot was 5.12. In this comparison only inverse linear order was significantly worse for ChatGPT 0-shot compared to the human participants with p = 0.0095. None of the other categories of rules were close to being significantly worse than the human participants with noun precedes adjective, passive conjugation, sentence final negation

and verb direct object swap receiving a p value of 0.9991, 0.9831, 0.9985 and 0.9859 respectively.

ChatGPT 1-shot was the final model we compared with the human participant benchmark, and this one-way ANOVA resulted in a F(db) value of 3.94. The pairwise comparison showed that none of the rule brackets were significantly different from the benchmark, with inverse linear order being the closest with p = 0.1029. Noun preceded adjective and passive conjugation received a p value of 0.9922 and 1.0 each. Finally sentence final negation and verb direct object swap had p = 0.9762 and 0.9669 respectively.

We also wanted to see how the models compared to each other, and we decided to start with NLLB200 and ChatGPT 0-shot. The one-way ANOVA resulted in F(db) = 11.40. Inverse linear order is the only category with a significant difference, with NLLB200 being significantly worse with p = 0.0218. The closest rule brackets to be significantly worse for NLLB200 were verb direct object swap and noun precedes adjective with p = 0.3617 and 0.5081. Passive conjugation and sentence final negation received a p value of 1.0 each.

Comparing NLLB200 to ChatGPT 1-shot gave the F(df) output of 12.46. This pairwise comparison showed that both inverse linear order and noun precedes adjective were significantly different, with NLLB200 being worse with p = 0.0011 and 0.0122 respectively. Passive conjugation, sentence final negation and verb direct object swap were not significantly different, with the comparison showing p = 0.9878, 0.9998 and 0.3883 respectively.

Finally we compared the two ChatGPT models against each other, with the one-way ANOVA resulting in F(db) = 5.02. This pairwise comparison showed no significant difference between the two models, with noun precedes adjective being the closest to being significantly different with p = 0.8683. In this comparison, both sentence final negation and verb direct object swap showed p = 1.0, while passive conjugation and inverse linear order received a p value of 0.998 and 0.9986 respectively.

5.8 Discussion

Although there is much debate within the machine learning community, universal grammar remains a contentious topic, with many arguing against its validity, certain models trained on high-resource languages can apply the knowledge gained from these languages to tasks involving languages for which they were not originally trained on. We believe this is only possible because of the principles of universal grammar the model learns from the high resource languages the model picks up along the way. To find proof of universal grammar, we conducted a series of translation experiments with different machine learning models and human participants with sentences manipulated with our very own universal grammar-defying rules. We set out with a goal of finding evidence for universal grammar by doing machine learning experiments, with the hypothesis that machine translation in low resource languages is made possible by regularities inherent in universal grammar.

From our research we found that human participants managed to achieve a good score on all five universal grammar-defying rule brackets we created, which we expected after reading about Musso et al.s findings in their paper on Brocas area. Using

the human results as a benchmark for the three machine learning model experiments, we found to our surprise that both ChatGPT 1-shot and 0-shot performed very well on all rule brackets, with the only significantly worse performance compared to the human participants coming from the 0-shot model on the sentences warped by our inverse linear order rule. We also discovered that NLLB200 performed significantly worse compared to our benchmark in the inverse linear order and verb direct object swap rule brackets. While surprising that NLLB200 performed so well on the sentence final negation, passive conjugation and noun precedes adjective rule brackets, the results from inverse linear order and verb direct object swap is more in line with what we expected to find when searching for evidence for universal grammar by doing machine learning experiments.

While surprising that none of the models performed significantly worse on the translation tasks with the rules noun proceeds adjective, passive conjugation and sentence final negation compared to the human participants, a more unexpected revelation was that the ChatGPT 1-shot model performed statistically equally to the human participants according to the pairwise comparison. We believe the reason why NLLB200 perform so much worse than the human participants and both ChatGPT models is because ChatGPT excels at pattern recognition, while the human mind have better capabilities when it comes to understanding the input provided. The human mind is obviously remarkably adept at solving problems based on perception, pattern recognition, memory and critical thinking. ChatGPT excels at pattern recognition and predicting next word, being able to capture relationships between words and the context in which they appear, allowing it to grasp nuances and subtle meanings. While pure translation models, like NLLB200, are specifically designed for translation text from one language to another. NLLB200 simply takes an input in one language and produces an equivalent output in another language. Models like NLLB200 are particularly useful for applications such as language translation services, where the goal is to convert text from one language to another while preserving the meaning, while ChatGPT can be used for chat-based applications, answering questions, providing information, and engaging in open-ended conversations. While NLLB200 tries to translate an ungrammatical Nynorsk sentence to English, ChatGPT will try to find patterns in the input provided instead. This is also reflected in the results of our ChatGPT 1-shot experiment, with six out of ten of the regular sentences in the inverse linear order bracket scored 1.0, while the absurd ones, aimed to prevent machine learning models from deducing context based solely on word recognition, only achieved two perfect 1.0 scores while NLLB200s scores were more evenly bad. NLLB200 also reportedly performs better with longer sentences, which also could contribute to the poor results, with our sentences having at most fourteen words.

While these results dont prove the existence of universal grammar, our findings provide some evidence in favor of the hypotheses, suggesting that good performance on low resource languages the model is not trained on is only possible because of the universal grammar the machine learning models learns from high resource languages. While the models we experimented with did perform above expectations on the sentences warped by the rules affecting the linear order of the words in the sentence less than other, sentence final negation and passive conjugation, NLLB200s poor performance on sentences warped by inverse linear order, verb direct object swap and to some degree noun precedes adjective aligns with our hypothesis that the machine

learning model would perform badly on translation tasks including universal-grammar defying sentences. ChatGPT 0-shots result on the rule altering the inverse linear order of the words in a sentence the most, inverse linear order, also supports our hypothesis to some degree.

It's conceivable that these results are attributable to a number of factors, including the nature of our original sentences short length and the use of the nllb-200-distilled-600M variant of NLLB instead of the larger nllb-200-3.3B version due to computational limitations. There is also a likelihood that one single human participant translating all one hundred sentences instead of ten participants translating ten of the same sentences would either improve or worsen the mean score of the human participant benchmark, in addition to making it more viable to compare with the machine learning models.

Chapter 6

Conclusions and Future Work

Expanding upon our research, future work could improve by working with NLLB200s 3.3B version instead of the 600M version used in this thesis. Creating more advanced and longer sentences before warping them with universal grammar-defying rules could also be done to get a fairer representation between ChatGPT and NLLB200. It would also be interesting to see if participants introduced to all one hundred sentences would increase or decrease the mean score of the human benchmark.

In conclusion, our exploration into the relationship between machine learning models, universal grammar, and translation tasks has yielded intriguing insights. While the rejection of the concept of universal grammar is prevalent among machine learning practitioners, our experiments suggest that certain models, particularly those trained on high-resource languages, leverage the principles of universal grammar to perform well on translation tasks involving languages for which they were not originally trained. Our primary goal was to uncover evidence for universal grammar by subjecting both machine learning models and human participants to translation experiments involving sentences manipulated by universal grammar-defying rules.

Remarkably, human participants exhibited a strong proficiency in translating sentences subjected to our universal grammar-defying rules, aligning with our expectations and drawing parallels to findings in the literature, such as Musso et al.'s work on Broca's area. Surprisingly, both ChatGPT 1-shot and 0-shot models demonstrated impressive performance across all rule brackets, with only a notable decline observed in the 0-shot model when confronted with sentences warped by our inverse linear order rule. Conversely, NLLB200 exhibited significantly poorer performance, particularly in the inverse linear order and verb direct object swap rule brackets, reinforcing our hypothesis that models like NLLB200 may struggle with translation tasks involving universal-grammar defying sentences.

The unexpected revelation that ChatGPT 1-shot performed statistically equally to human participants on certain rule brackets, coupled with its aptitude for capturing relationships and nuances, emphasizes the model's proficiency in pattern recognition. In contrast, NLLB200's struggle may be attributed to its focus on literal translation without the depth of pattern recognition exhibited by human participants and ChatGPT.

Although our results do not conclusively prove the existence of universal grammar, they provide compelling evidence supporting our hypothesis. The alignment of machine learning model performance with human participants, especially in the face of universal grammar-defying rules, hints at the role of universal grammar in facilitating

effective translation tasks. Factors such as sentence length and the choice of NLLB variant may have influenced our findings, highlighting avenues for future research to refine and expand upon our experimental framework.

In essence, our study contributes valuable insights into the interplay between machine learning models, universal grammar, and translation tasks, shedding light on the potential influence of universal grammar principles on the success of these models in handling diverse linguistic challenges.

Bibliography

Evans, N., and S. Levinson (2009), The myth of language universals: Language diversity and its importance for cognitive science, *The Behavioral and brain sciences*, *32*, 429–48; discussion 448, doi:10.1017/S0140525X0999094X. 1.1

Meta (2022), 200 languages within a single ai model: A breakthrough in high-quality machine translation, https://ai.meta.com/blog/nllb-200-high-quality-machine-translation/, accessed: (30.05.2023). 2.2

Musso, M., A. Moro, V. Glauche, M. Rijntjes, C. Büchel, and C. Weiller (2003), Broca's area and the language instinct, *Nature neuroscience*, *6*, 774–81, doi:10.1038/nn1077. (document), 1.1, 2.1

NLLB-Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. M. Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang (2022), No language left behind: Scaling human-centered machine translation. 2.2

Techvify-Software (2023), Gpt-3.5 vs gpt-4: Exploring unique ai capabilities, https://techvify-software.com/gpt-3-5-vs-gpt-4/, accessed: (01.11.2023). 2.3

Wu, S., and M. Dredze (2020), Are all languages created equal in multilingual bert?, pp. 120–130, doi:10.18653/v1/2020.repl4nlp-1.16. (document)

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, ukasz Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, J. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean (2016), Google's neural machine translation system: Bridging the gap between human and machine translation. 5.1