Evaluation of Machine Translation: Greek to English

Stefanos Sfinarolakis (inf2021218)
Nikolaos Trypakis (inf2021229)
Konstantinos Kafteranis (inf2021090)

Abstract

Machine Translation (MT) has made significant strides, yet languages like Greek, with intricate syntactic structures, remain underexplored. This study investigates the challenges faced by MT systems in translating Greek to English, specifically focusing on negation, coordination, subjunction, punctuation, and metaphors. Using two large language models (LLMs), Claude and Microsoft Copilot, we evaluate their translation accuracy across 1,500 Greek sentences. Our findings reveal that Copilot excels with basic sentence structures, while Claude performs better with complex contexts. Through BLEU score analysis and error assessment, we highlight the difficulties in handling idiomatic expressions and scientific terminology. Further research is needed to improve LLM performance for low-resource languages and specialized linguistic features.

1. Introduction

Machine Translation has become a widely used tool in the realm of natural language processing (NLP), making communication easier and bridging the linguistic gaps among languages. MT systems have shown significant progress, but their best performance is also seen in the most famous and widely spoken languages, such as English, Spanish, or Chinese.

Less common languages like Greek are underrepresented, making even state-of-the-art tools struggle with Greek-to-English translation sometimes. Greek is a complex language with lots of grammar rules and flexible sentence structures, which makes it tricky for machine translation (MT) systems. Unlike English, Greek uses things like flexible word order, detailed verb forms, and case markers to show meaning. These features can be hard to translate into English, which follows stricter grammar rules. This often leads to mistakes or loss of meaning in translations. Another big challenge is the lack of good-quality, large datasets for translating Greek to English. Without enough data, it's hard to train MT systems to handle the unique features of Greek. As a result, Greek is often overlooked in MT research, with most efforts focused on bigger, more commonly used languages.

In this study, we aim to assess the effectiveness of Greek-to-English MT by reviewing existing literature and identifying the key challenges discussed in previous research. We then focus on specific grammatical phenomena of interest and compile a dataset to address these aspects. Furthermore, we conduct our own experiments using two different large language models (LLMs) to explore their capabilities and limitations in translating between Greek and English. By comparing their performance, we gain insights into the strengths and weaknesses of these models for this specific language pair.

2. Methodology

In this section, we outline the steps taken in our research. First, we review existing literature on Greek-to-English Machine Translation (MT) to identify key challenges and approaches. Then, we describe our experimental methodology using two large language models (LLMs), Claude and Mi-

crosoft Copilot. Finally, we introduce the dataset used in our experiments, explain its linguistic structure, and detail the evaluation metrics and error analysis approach.

2.1. Overview of Large Language Models

To evaluate the performance of MT for Greek-to-English translation, we employ two prominent LLMs:

- **Claude**: Claude is a state-of-the-art language model designed for a range of natural language processing (NLP) tasks. It is particularly noted for its conversational abilities and its capacity to process nuanced linguistic phenomena. Claude utilizes advanced transformer-based architectures and has been trained on diverse datasets, enabling it to handle complex sentence structures.
- **Microsoft Copilot**: Microsoft Copilot is an AI assistant that integrates advanced NLP capabilities into various applications. It leverages models from the OpenAI ecosystem, such as GPT, to perform translation tasks. While not explicitly designed for MT, Copilot demonstrates strong contextual understanding and adaptability across languages.

2.2. Dataset

The dataset used in our experiments consists of 1,500 sentences from here, divided into five linguistic phenomena encountered in Greek-to-English translation, that we choose. Our criteria for phenomena selection has to do with our assumption based on the bibliogrpahy that most difficult thing to translate are phrases with special caltural meaning. The rest were chosen based on the easy identification throught regular expressions using a script to detect with a script we created. These phenomena along the number of sentances are described as follows:

- **Negation (629)**: Negation in Greek often involves multiple words or flexible syntax, posing challenges for accurate translation into English, which generally uses fixed patterns for negation.
- **Coordination (401)**: Greek's coordination structures, which often allow ellipsis or flexible conjunctions, may result in ambiguities when translated into English.
- **Subjunction (138)**: Subjunction involves subordinate clauses, which in Greek can exhibit diverse word orders and conjunctions that do not always directly map to English equivalents.
- **Punctuation (332)**: Greek punctuation rules differ slightly from English, with variations in the usage of commas and quotation marks, affecting sentence segmentation during translation.
- **Metaphors and Special Phrases (10)**: Greek frequently employs metaphors and idiomatic expressions that require cultural or contextual understanding for accurate translation into English.

2.3. Evaluation and Error Analysis

The translation quality is evaluated using the BLEU (Bilingual Evaluation Understudy) metric. The process includes the following steps:

- Translations are generated using both Claude and Microsoft Copilot for each linguistic phenomenon.
- BLEU scores are calculated for each phenomenon individually and for the entire dataset as a whole.
- Errors are analyzed to identify patterns, such as specific mistakes made by each model.

This methodology allows us to assess the relative strengths and weaknesses of the two LLMs .

3. Literature Overview

The development of Neural Machine Translation (NMT) models based on the Transformer architecture has shown significant progress in handling Greek and English translations. These models excel at tasks such as translating Greek text, extracting triplets, and retranslating back to Greek, outperforming current methods for the Greek language. However, challenges remain in accurately interpreting culturally charged expressions, highlighting the complexity of reconciling linguistic and cultural nuances [4]. Early systems such as METIS and METIS-II used statistical data and monolingual corpora without bilingual resources for Greek-English translation, while Uplug, a word-alignment system, contributed to the creation and evaluation of a reliable Greek-English bilingual dictionary. Research has also compared the performance of KantanMT and Moses statistical translation systems within a cross-lingual information retrieval (CLIR) framework, noting the superior average accuracy of KantanMT despite the morphological complexity of Greek [4]; [2].

In addition, multilingual models often have limitations due to their predominant training on English data, resulting in reduced accuracy for low-resource languages. Reliance on low-quality or automated translations contributes to errors, while the Anglophone bias of these systems risks distorting local cultural meanings and making inappropriate associations. These challenges extend to other domains, as demonstrated by a multilingual mental health dataset developed for severity prediction in languages such as Turkish, French, Portuguese, German, Greek, and Finnish. The dataset revealed significant prediction variability and lower accuracy in low-resource languages due to cultural nuances, limited data, inconsistent translations, and inadequate error correction. These findings highlight the need to use Large Language Models (LLMs) as assistive rather than autonomous tools for mental health professionals, paralleling broader multilingual NLP issues [5]; [2].

In addition, structural, semantic, and syntactic divergences between languages add further complexity to the task of aligning language representations. These divergences, especially between morphologically rich and resource-constrained languages, pose significant challenges for multilingual models. The inability to fully capture cultural and contextual subtleties often results in degraded translation quality and biased associations, underscoring the urgent need for novel approaches to address linguistic differences and improve model effectiveness [3].

The lack of sufficient bilingual resources and the reliance on monolingual corpora also exacerbate these challenges. Studies show that low-resource language translations in particular suffer from significant context loss, which affects downstream applications. In addition, the design and evaluation of language-specific pre-training schemes to overcome these hurdles are highlighted as critical steps forward. They emphasize the need to address the lack of domain-specific vocabularies and the over-reliance on generalized linguistic data, which hinder the performance of linguistic models in various applications, including minority language scenarios [1]

4. Experimental Results

4.1. Numeric Results

The numeric results are shown in tables 1 and 2

4.2. Limitations

This approach has several limitations:

- **Dataset Size:** The dataset is small, with only 1,500 sentences, limiting the diversity and complexity of linguistic phenomena tested.
- **Simple Linguistic Phenomena:** Most of the phenomena, such as negation and coordination, are generally easy for modern LLMs to handle.
- **Limited Metaphors and Special Phrases:** The dataset included only 10 sentences for metaphors and special phrases, which are the most challenging category. This small sample size makes it difficult to fully assess model performance, even though many cases were handled reasonably well.
- **Limitations of BLEU Metric:**
  - BLEU penalizes minor differences, such as "cannot" vs. "can't," even when the meaning is correct.
  - It evaluates full sentences, where a small mistake can significantly lower the score.
  - BLEU may fail to reward valid translations that differ in phrasing from the reference.

These limitations suggest that while the study provides useful insights, it does not fully capture the models' strengths, especially for nuanced translation tasks.

5. Conclusions and Proposals

Nowadays, MT models are advanced and generally perform really well on translating individual sentences, however some problems may still occur. After testing the five phenomenons we mentioned before, it seems that both models manage to identify almost correctly punctuation (Copilot: 59.97, Claude: 59.79) and negation (57.14, 60.55). For coordination, Copilot seems to suffer (21.21) while Claude manages to get double the score (47.05). Similarly, Copilot struggles with subjunctions (26.54) while Claude looks more confident (49.38). The failure to recognise these two phenomenons can lead to different syntactical linking of the words within the sentence making it difficult to express its original meaning. However, this was a rare case and the low score is due to Bleu's limitations on interpreting the sentences rather than the models themselves. As expected and understood from our bibliography, both Copilot and Claude appear to have a problem with metaphors and phrases (27.69, 40.21). They tend to translate them literally instead of understanding the intended meaning. However, Claude seems to perform quite better than Copilot in terms of handling complex phrases, as it is more likely to choose accurate technical

Grammatical Phenomenon	Number of Sentences	Score (Copilot)	Score (Claude)
Negation	629	57.14	60.55
Coordination	401	21.21	47.05
Subjunctions	138	26.54	49.38
Punctuation	332	59.97	59.79
Metaphor/Special Phrase	10	27.69	40.21

Πίνακας 1: Linguistic phenomena analysis: number of sentences and BLEU scores by Microsoft Copilot and Claude.

terms. Most mistakes, aside from special phrases, were random false recognition of punctuation symbols and change of words. Those cases were not common and don't indicate any significant pattern. One last thing of significance is the gender misinterpretation which makes sense since without the right context the models has to just take a guess. Lastly, we would like the suggest as a continuation of this research to use bigger dataset of special phrases and AI models specifically trained on greek to english instead these widely used LLMs presented in this work.

6. Refernces

έ

[1] Omkar Khade .ά. "Challenges in Adapting Multilingual LLMs to Low-Resource Languages using LoRA PEFT Tuning". *Pune Institute of Computer Technology* (2024). Indian Institute of Technology Madras, L3Cube Labs Pune.

[2] Ourania Kolovou. "Machine Translation from Ancient Greek to English: Experiments with OpenNMT". /ή. Gothenburg, Sweden: University of Gothenburg, 2023. URL: https://gupea.ub.gu.se/bitstream/handle/2077/81765/Machine%20Translation%20from%20Ancient%20Greek%20to%20English_Experiments%20with%20OpenNMT.pdf?sequence=1&isAllowed=y.

[3] Gabriel Nicholas Aliya Bhatia. *Lost in Translation: Large Language Models in Non-English Content Analysis*. https://examplelink.com. Accessed: 2025-01-20. ά. 2023.

[4] Dimitris Papadopoulos .ά. "PENELOPIE: Enabling Open Information Extraction for the Greek Language through Machine Translation". *Proceedings of the EACL 2021 SRW* (2021). arXiv: 2103.15075. URL: https://arxiv.org/abs/2103.15075.

[5] K. Skianis, J. Pavlopoulos A. S. Doğruöz. "Severity Prediction in Mental Health: LLM-based Creation, Analysis, Evaluation of a Novel Multilingual Dataset". (2024). URL: https://arxiv.org/pdf/2409.17397.

Πίνακας 2: Error Analysis of Translations

Original (Greek)	Expected Translation	Copilot Translation	Claude Translation	Comments
Κάλλιο πέντε και στο χέρι, παρά δέκα και καρτέρει.	A bird in the hand is worth two in the bush.	Better five in hand than ten waiting.	Better five in hand than ten waiting.	The models translated word by word, failing to understand the metaphorical meaning of the phrase.
Ανακοινώθηκε πως η αιτία του θανάτου ήταν ο καρκίνος των ενδοηπατικών χοληφόρων οδών.	The cause of death was announced as intrahepatic bile duct cancer.	It was announced that the cause of death was cancer of the intrahepatic bile ducts.	It was announced that the cause of death was cholangiocarcinoma.	The models struggled with the subjunction structure, causing variations in how the disease was named.
Ανεξαρτησία· Δεν ενδιαφέρομαι.	Independence? I'm not interested.	Independent? I'm not interested.	Independent? I'm not interested.	The models translated "ί" as "Independent" instead of "Independence," which changed the meaning of the sentence.
Γύρισε το χερούλι για να ανοίξει την πόρτα.	She turned the handle to open the door.	Turn the handle to open the door.	Turn the handle to open the door.	The models missed the subject "She," giving a more general translation.
Δεν απάντησε.	She didn't answer.	He didn't answer.	He didn't answer.	The models incorrectly translated the subject "She" as "He."
Ξέρουν ότι είσαι εδώ·	Do they know you're here?	They know you are here.	Do they know you are here?	The machines missed the tone of the Greek sentence, which is important for the correct translation.
Ξεχωρίζω σαν την μύγα μες το γάλα.	I stick out like a sore thumb.	I stand out like a sore thumb.	I stand out like a fly in milk.	One model incorrectly translated the phrase word by word, while the other used the appropriate English idiom.
Ο Τομ είναι γυμνός από τη μέση και πάνω.	Tom is naked from the waist up.	Tom is naked from the waist up.	Tom is shirtless.	Claude used "shirtless" instead of translating the phrase directly.
Έχει μάτια και στην πλάτη.	She has eagle eyes.	He has eyes in the back of his head.	He has eyes in the back of his head.	The machines translated word by word instead of understanding the meaning of the phrase.
Κάποια άτομα έχουν μη σταθερούς πυρήνες και αυτό τα οδηγεί στη διάσπαση αν τους ασκηθεί μικρή ή ακόμα και καθόλου πίεση.	Some atoms have unstable nuclei which means that they tend to break apart with little or no nudging.	Some individuals have unstable nuclei, leading them to decay under little or no pressure.	Some people have unstable cores that lead them to disintegration if even little or no pressure is applied.	The models mistakenly translated "atoms" as "people" or "individuals" instead of understanding the scientific context.
Κάποιου του χάριζαν ένα γάιδαρο και τον κοίταζε στα δόντια.	Don't look a gift horse in the mouth.	Someone was given a donkey as a gift, and they were looking at its teeth.	Someone was being given a donkey and was looking it in the mouth.	The machines translated word by word instead of understanding the meaning of the phrase.