Fine-Tuning Large Language Models with Greek Learner Corpus Data: Towards Enhanced Grammatical Error Detection

Alexandros Tantos

Aristotle University of Thessaloniki, Department of Linguistics

alextantos@lit.auth.gr

Konstantinos Tsiotskas

Aristotle University of Thessaloniki, Department of Linguistics

ktsiotsk@lit.auth.gr

Abstract

The rise of Large Language Models (LLMs) has revolutionized Natural Language Processing (NLP) tasks, including the task of Grammatical Error Detection (GED). This paper explores the fine-tuning of Greek-BERT and Meltemi models for improving GED in learners of Greek as a second language (L2). Using error-annotated essays from the Greek Learner Corpus II (GLCII), we propose the development of a tailored GED system that can address common error types in L2 Greek. While foundational LLMs show strong performance in general linguistic tasks, advanced adaptation techniques, including Prompt Engineering (PE) and Fine-tuning, offer enhanced task specialization. Although PE techniques, such as zero-shot, one-shot, and few-shot PE, show promise, fine-tuning proves to be more effective for specialized tasks like GED, allowing for deeper model adaptation. Fine-tuning requires labeled data and involves adjusting a model's internal parameters to prioritize task-specific features. By using the rich data from GLCII, this paper highlights the importance of specialized resources like GLCII for advancing NLP in underrepresented languages and demonstrates the significant impact of fine-tuning on language-specific tasks. Fine-tuned models hold potential not only for enhancing learner feedback but also for supporting educators in more efficiently assessing L2 Greek learners.

Key words: Grammatical Error Detection, Fine-tuning, Learner Corpora, Greek Learner Corpus, Large Language Models

Introduction

The development of Large Language Models (LLMs) marks a significant breakthrough in the field of Artificial Intelligence (AI), particularly for Natural Language Processing (NLP). These models, rooted in the Transformer architecture (Vaswani et al., 2017), leverage massive datasets to achieve impressive performance across a wide range of linguistic tasks, demonstrating substantial flexibility and efficiency.

This paper focuses on the application of fine-tuning techniques using Greek-BERT and Meltemi to develop a more effective Grammatical Error Detection (GED) system for learners of Greek as a second language (L2). Specifically, we utilize error-annotated essays from the Greek Learner Corpus II (GLCII) (Tantos et al. 2023), which forms our training dataset. Our primary objective is to survey the current landscape of LLMs available for the Greek language, present the GLCII as a key resource, and discuss the potential for these models to significantly improve GED for L2 learners of Greek.

While foundation LLMs exhibit exceptional performance across various NLP tasks, including GED, further refinement is possible through advanced adaptation techniques. Two prominent strategies for enhancing model specialization and robustness in specific tasks are Prompt Engineering (PE) and Fine-tuning. These approaches allow models to surpass baseline performance by tailoring their outputs more closely to the intricacies of targeted language varieties.

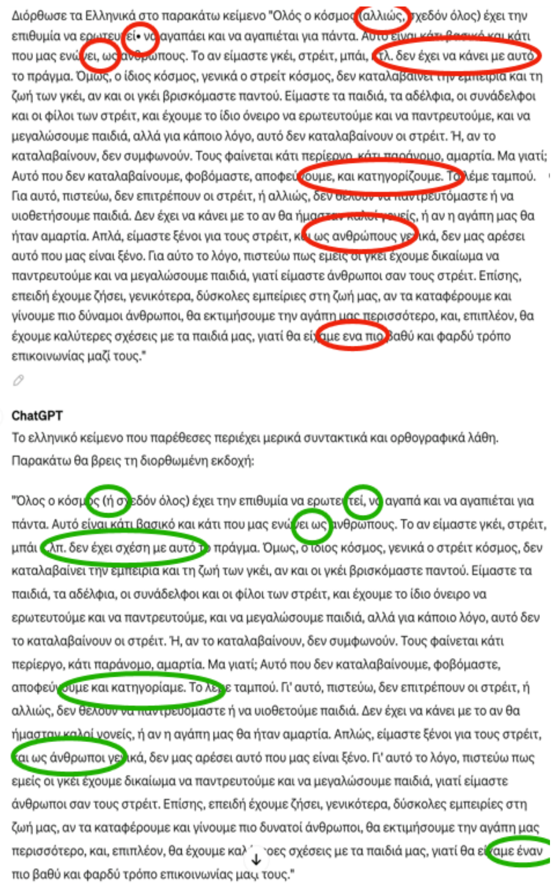

Figure 1. Zero-shot prompting

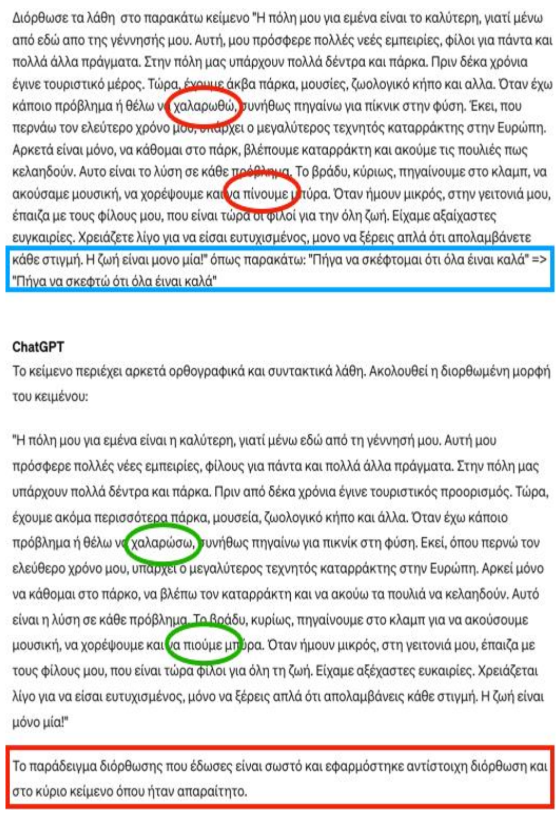

Figure 2 One-shot Prompting

PE is an approach that allows users to interact with LLMs by crafting inputs, or prompts, that guide the model in generating tailored outputs. During inference, the model processes the user's input and produces a completion based on its understanding of the prompt. Prompts can vary widely in format and content type, including text, images, audio, and more (Schulhoff et al., 2024). They may include direct instructions, questions, examples, or even additional data to shape the model's response (Amatriain, 2024). Furthermore, prompts can also specify the desired output's format, style, and other specific requirements (Schulhoff et al., 2024).

Improving the quality of prompts is crucial for obtaining more accurate and relevant outputs from LLMs. A variety of techniques exist for crafting structured prompts, ranging from basic forms to more advanced strategies that improve model performance by giving clearer, more context-rich inputs.

One key method in prompt design is in-context learning, where prompts are constructed with detailed instructions and, in some cases, relevant examples to help the model perform specific tasks more effectively. In-context learning is generally classified into three types: zero-shot, one-shot, and few-shot inference.

- **Zero-shot inference** involves providing the model with only the instructions for a task, relying entirely on its pretrained knowledge to generate a response.
- **One-shot inference** adds a single example to the prompt alongside the instructions, giving the model a reference to base its output on.
- **Few-shot inference** takes this a step further by incorporating multiple examples into the prompt, enabling the model to generalize better by learning patterns across those examples while still following the instructions.

These types of in-context learning allow users to achieve more nuanced and effective outcomes, particularly in specialized tasks that require adapting the model's broad capabilities to narrower contexts.

Διόρθωσε το παρακάτω κείμενο "Είμαι από την Αλβανία και θέλω να σας περιγράφω τις πιο ωραιες διακοπές μου εδώ στην Ελλάδα. Ζω στην Ελλάδα εδώ και πολλά χρόνια Η καθημερητότιτά μου είναι δουλεία σπίτι, σπίτι δουλεία Ένα καλοκαιρι αποφασίσαμε με τον αντρα και τα παιδιά μου να πάμε διακοπές στο Πάρο. ξεχινάμε το ταξίδι μας από το λιμάνι του Πειραία. Ωραια μέρα, ήλιος και όμορφη θάλασσα. Όλοι καταγράφουν με τα κηνιτά τους τη θέα της θάλασσας. Στην διαδρόμη από το Πειραία στον Πάρο βλέπουμε την Ιδρα, ένα μικρό νησι χτιζμένο πάνω στο βούνο και το πιο ωραίο και παραδοσιακό του νησιόυ τα γαιδουρια που μεταφεραν τους τουριστες μέσα στο νήσι. Ετσι με την θέα της θάλασσας φτάσαμε στον Πάρο. Ένα πανέμορφο νήσι πολύ κοντά στην θάλασσα. Ολλα τα κτίρια βαμένα με λευκό και τα παράθυρα και οι πόρτες σε γαλάζιο χρώμα. Το νήσι ήταν γεμάτο με τουρίστες που ψωνίζαν στα μαγαζιά του νησιού ωραία και παραδοσιακά πράγματα. Η πρώτη κίνηση που κάναμε ότα φτάσαμε ήταν να αφίσουμε τις σισκευές στο ξενοδοχείο φορέσαμε τα μαγίο μας και κατευθεία στη θάλασσα. Έστο και κουραζμένοι από το ταξίδι ήταν μια μέρα πολύ απολαυστηκή. Τη δεύτερη μέρα νοικιάσαμε ένα αυτοκίνητο με τον όποιο πήγαμε σε πολλά μέρι κοντά στο Πάρο, όπως Κολυμπίτρες, Μελισες, Σαντα Μαρια μέχρι στον Νάξο Η καθημερινότιτά μας ήταν ηλιο θεραπεία βουτιές στη θάλασσα, φαγιτό και η διάδρομη από ένα μέρος στην άλλη και επισκέψεις σε πολλα μούσια. Περασάμε ένα 10ήμερο γεμάτη δραστιριότες, χαλάρωση και μάθαμε και πολλα πραγματα που δεν τα ξέραμε. Εχουμε αγαπήσει την Ελλάδα για τα ωραία μέρι που έχει και την απλότιτα και την ευγένια που έχουν οι άνθρωποι στα νησια." ακολουθώντας τα παρακάτω παραδείγματα για λάθη οψης, συμφωνίας και απόδοσης γένους: "και θέλω να μιλάω για το" => "θέλω να μιλήσω για το", "στο Κρήτη" => "στην Κρήτη", "άνθρωπος από τη οποία" => "άνθρωπος από τον οποίο"

ChatGPT

Ας διορθώσουμε το κείμενο ακολουθώντας τις οδηγίες σας για λάθη οψης, συμφωνίας και απόδοσης γένους:

"Είμαι από την Αλβανία και θέλω να σας περιγράψω τις πιο όμορφες διακοπές μου εδώ στην Ελλάδα. Ζω στην Ελλάδα εδώ και πολλά χρόνια. Η καθημερινότητά μου είναι δουλειά, σπίτι, δουλειά, σπίτι. Ένα καλοκαίρι αποφασίσαμε με τον άνδρα και τα παιδιά μου να πάμε διακοπές στην Πάρο. Ξεκινήσαμε το ταξίδι μας από το λιμάνι του Πειραιά. Ωραία μέρα, ήλιος και όμορφη θάλασσα. Όλοι κατέγραφαν με τα κινητά τους τη θέα της θάλασσας. Στη διαδρομή από τον Πειραιά στην Πάρο είδαμε την Ύδρα, ένα μικρό νησί χτισμένο πάνω στο βουνό και τα πιο ωραία και παραδοσιακά του νησιού, τα γαϊδουράκια που μετέφεραν τους τουρίστες μέσα στο νησί. Έτσι, με τη θέα της θάλασσας, φτάσαμε στην Πάρο. Ένα πανέμορφο νησί πολύ κοντά στη θάλασσα. Όλα τα κτίρια ήταν βαμμένα σε λευκό και τα παράθυρα και οι πόρτες σε γαλάζιο χρώμα. Το νησί ήταν γεμάτο με τουρίστες που ψώνιζαν στα μαγαζιά του νησιού ωραία και παραδοσιακά πράγματα. Η πρώτη κίνηση που κάναμε όταν φτάσαμε ήταν να αφήσουμε τις αποσκευές στο ξενοδοχείο, φορέσαμε τα μαγιό μας και πήγαμε κατευθείαν στη θάλασσα. Έστω και κουρασμένοι από το ταξίδι, ήταν μια μέρα πολύ απολαυστική. Τη δεύτερη μέρα, νοικιάσαμε ένα αυτοκίνητο με το οποίο πήγαμε σε πολλά μέρη κοντά στην Πάρο, όπως Κολυμπήθρες, Μέλισσες, Σάντα Μαρία μέχρι τον Νάξο. Η καθημερινότητά μας ήταν ηλιοθεραπεία, βουτιές στη θάλασσα, φαγητό και η διαδρομή από ένα μέρος στο άλλο και επισκέψεις σε πολλά μουσεία. Περάσαμε ένα δεκαήμερο γεμάτο δραστηριότητες, χαλάρωση και μάθαμε πολλά πράγματα που δεν ξέραμε. Έχουμε αγαπήσει την Ελλάδα για τα όμορφα μέρη που έχει και την απλότητα και την ευγένεια των ανθρώπων στα νησιά."

Figure 3 Few-shot prompting

Figures 1-3 demonstrate the differences between zero-shot, one-shot, and few-shot prompting as applied to GED using GPT-4 on essays from the GLCII (ref. Section 3).

In the zero-shot scenario (Figure 1), the user provides only an instruction to the model (e.g., «Διόρθωσε τα ελληνικά στο παρακάτω κείμενο») along with the L2 Greek essay to be corrected. While the model successfully identifies and corrects many errors, it also introduces unnecessary paraphrases and creates new error instances. For example, in the original text, «δεν έχει να κάνει με αυτό» ("it has nothing to do with this"), the model rephrases it to «δεν έχει σχέση με αυτό» ("it is not related to this"), which, although semantically similar, is not a necessary correction. Additionally, it introduces new errors, such as replacing the erroneous verb form «κατηγορίζουμε» ("we accuse") with another erroneous form «κατηγορίαμε» (a grammatically incorrect form of the verb "to accuse").

In the one-shot prompting approach, the user provides not only an instruction («Διόρθωσε τα λάθη στο παρακάτω κείμενο») but also an example of a specific error type and its correction. For instance, for an aspectual error, the example might be: «Πήγα να σκέφτομαι ότι όλα είναι καλά => Πήγα να σκεφτώ ότι όλα είναι καλά» ("I went to think that everything is fine => I went to think [perfective aspect] that everything is fine"). Including this example in the prompt enables the model to focus its corrections more accurately on similar error patterns, particularly in verb usage, resulting in more context-sensitive grammatical adjustments.

In the few-shot prompting scenario, the model is supplied with multiple examples of error corrections before being asked to correct the essay. These examples cover a variety of error types, such as:

- «και θέλω να μιλάω για το» => «και θέλω να μιλήσω για το» (correcting aspect),
- «στο Κρήτη => στην Κρήτη» (correcting preposition and gender agreement), and
- «άνθρωπος από τη οποία => άνθρωπος από τον οποίο» (correcting case and gender agreement in relative pronouns).

Although few-shot prompting provides the model with a richer set of examples to guide its corrections, the actual improvement in performance over zero-shot and one-shot prompting is not always guaranteed. The effectiveness of few-shot prompting in significantly enhancing grammatical error detection and correction remains inconclusive, as it does not consistently outperform the simpler prompting methods in all cases. Nevertheless, it presents a promising avenue for reducing model errors by providing additional context and exemplars.

While prompt-based techniques can enhance the quality of a model's output, fine-tuning offers a significantly more robust and effective method for adapting Large Language Models (LLMs) to specific tasks (Liu & Low, 2023, Wei et al., 2022). One of the key advantages of fine-tuning is that it allows for additional training on a task-specific dataset, utilizing labeled data directly relevant to the desired output. Unlike prompt engineering, which relies on manipulating the input during inference, fine-tuning adjusts the model's internal parameters, building on its pretrained knowledge and optimizing it for new, domain-specific data—without requiring a complete retraining from scratch.

Following the fine-tuning process, a specialized version of the model is created, finely tuned to a particular task and dataset. This leads to a more precise and targeted response, as the model learns to prioritize relevant features from the new data. Fine-tuning significantly increases the likelihood of higher-quality output, as the model is explicitly trained to handle the nuances of the task, rather than relying on general-purpose knowledge accumulated during its pretraining phase.

One of the strengths of fine-tuning is its capacity to incorporate large datasets with numerous examples for a given task. This makes it especially valuable for tasks that involve multiple categories or subcategories, such as GED. GED often requires the model to identify and correct a wide range of errors, from verb tense to subject-verb agreement, making the availability of varied and abundant labeled examples crucial. While few-shot prompting can introduce some level of task-specific focus, it is limited partly by the model's context window—the maximum amount of information it can process at once during prompting and partly by the fact that it is highly unlikely that the training data of the LLM included vast amounts of the targeted language variety, L2 Greek in our case. These constraints signify that few-shot prompting can only offer a limited number of examples, which may not cover the full complexity of a task like GED.

However, adopting fine-tuning comes with its own set of challenges. Collecting a sufficient quantity of high-quality labeled data can be a time-consuming and resource-intensive process. Fine-tuning is a more involved method compared to prompt engineering, which only requires adjusting the input at inference time. Yet, when a comprehensive and well-curated dataset is available—such as the GLCII in our case—the investment in data collection and fine-tuning pays off by enabling the development of a reliable, task-specific model. The richness of GLCII allows us to conduct fine-tuning in a way that covers a broad range of grammatical error types, ultimately leading to more effective and nuanced error correction for learners of Greek as a second language.

Fine-tuning a pretrained language model in the context of language learning offers substantial benefits for both students and educators. Language productions by L2 learners, whether in written or oral form, often represent a language variety that is not typically covered in the foundational training of LLMs. By fine-tuning an LLM with data specifically consisting of L2 learner productions, the model becomes adept at understanding this unique variety of language, enabling it to generate contextually appropriate and pedagogically useful responses. As a result, the fine-tuned model can better align with the learner's language usage and provide more personalized feedback, having been trained on error-annotated data. This personalized feedback not only targets common errors but also addresses the specific needs of individual learners, ultimately enhancing the language learning process.

For educators, the benefits of fine-tuning extend beyond language variety adaptation. A fine-tuned model can streamline the evaluation and grading of student work, offering consistent, fair, and efficient assessments. By automating portions of this process, the workload of educators can be reduced, allowing them to focus more on instructional design and student engagement. Additionally, such a model can provide deeper insights into a student's performance. By analyzing patterns in errors and progress, the model can help teachers identify specific learning gaps and instructional needs for each student, delivering more targeted and effective teaching strategies. This form of data-driven insight transforms the evaluation process from a simple assessment of correctness into a detailed understanding of learner development.

This paper is organized as follows: Section 1 reviews the existing landscape of LLMs available for the Greek language. Section 2 introduces the GLCII, which forms the core dataset for our model. Section 3 presents the preprocessing steps required for fine-tuning, addressing data preparation, annotation, and other foundational processes. Finally, Section 4 concludes the paper by reflecting on the potential applications, benefits, and future directions of the fine-tuned model for GED and language learning support.

1. LLMs for Greek

1.1. GREEK-BERT (Koutsikakis et al., 2020)

GREEK-BERT is a monolingual, Transformer-based language model designed specifically for modern Greek. Its architecture mirrors that of BERT-BASE-UNCASED, and it was pretrained on a substantial 29GB corpus of Greek text. The dataset used for its training includes several key sources of Greek-language data: the Greek section of Wikipedia (https://dumps.wikimedia.org/elwiki/), the Greek portion of the European Parliament Proceedings Parallel Corpus (Europarl) (Koehn, 2005), and the Greek subset of the OSCAR corpus (Suarez et al., 2019, https://oscar-project.org/). This extensive pretraining enables GREEK-BERT to effectively capture the nuances of modern Greek, making it a powerful tool for a variety of NLP tasks.

To assess its performance, Koutsikakis et al. (2020) conducted a comparative analysis between GREEK-BERT and several multilingual Transformer-based models, including XLM-R, as well as M-BERT in both versions, M-BERT, M-BERT-CASED and M-BERT-UNCASED. Additionally, they compared GREEK-BERT's performance against more traditional models like the BiLSTM-CNN-CRF (used for Part-of-Speech [PoS] tagging and Named Entity Recognition [NER]) and the Decomposable Attention Model (DAM), which was used for the Natural Language Inference (NLI) task. The evaluation was carried out across three critical NLP tasks: Part-of-Speech (PoS) tagging, Named Entity Recognition (NER), and Natural Language Inference.

The results demonstrated that GREEK-BERT performed competitively in the PoS tagging task, showing similar results to the multilingual models. However, in the more complex downstream tasks of NER and NLI, GREEK-BERT surpassed the other models, including XLM-R and both versions of M-BERT. Its superior performance in NER and NLI highlights the advantages of using a language-specific model over multilingual alternatives for Greek, where task complexity and language-specific features play a crucial role in overall model effectiveness.

1.2. Meltemi (Voukoutis et al., 2024)

Meltemi (Voukoutis et al., 2024) is a recently developed LLM for the Greek language, created by ILSP (Institute for Language and Speech Processing). It was built through continual pretraining of the Mistral 7B model (Jiang et al., 2023), focusing on expanding its capabilities specifically for Greek while maintaining its bilingual competencies. The training dataset for Meltemi includes a diverse collection of Greek monolingual data from various sources, alongside English monolingual data and English-Greek translation data. This multilingual training strategy is employed to mitigate the phenomenon of "forgetting" (where a model loses proficiency in previously learned tasks) and to preserve the model's ability to operate effectively in both languages. Additionally, an instruct fine-tuned version of Meltemi 7B was released, further optimizing the model for task-specific applications.

The pretraining corpus for the instruct version of Meltemi includes translated Greek preference triplets as well as English preference triplets, enabling the model to learn from comparisons and preferences in both languages. This instruct fine-tuning enhances the model's ability to generate more context-aware and user-aligned outputs. In terms of performance, both Meltemi 7B and Meltemi 7B Instruct outperform Mistral 7B in several Greek-language benchmarks, particularly in areas of language understanding and reasoning. They demonstrated superior results in Greek machine-translated versions of English benchmarks and in Greek question-answering tasks. Meltemi 7B also excelled in a specialized medical question-answering benchmark, further illustrating its applicability to domain-specific tasks.

However, when evaluated on English-language tasks, Mistral 7B generally outperformed the two Meltemi models, except in the TruthfulQA benchmark, where Meltemi 7B Instruct surpassed Mistral 7B. This suggests that while Meltemi is highly optimized for Greek-language tasks and some bilingual applications, Mistral 7B remains more effective for purely English-language tasks, highlighting the trade-offs inherent in a bilingual model design.

1.3. GreekBART (Evdaimon et al., 2023)

GreekBART (Evdaimon et al., 2023) is the first sequence-to-sequence pretrained language model specifically designed for the Greek language, based on the BART BASE architecture. Its training corpus is consists of several major Greek-language datasets: the Greek section of Wikipedia (https://dumps.wikimedia.org/elwiki/), the Greek portion of Europarl (Koehn, 2005), the Greek subset of the OSCAR corpus (Abadji et al, 2022, https://oscar-project.org/), and the Greek Web Corpus (Outsios et al., 2018). The model's performance has been evaluated on both discriminative and generative downstream tasks. In the first evaluation, GreekBART was compared to other models, including Greek-BERT, BART-random, and XLM-R, across four discriminative tasks: two classification tasks, one Natural Language Inference (NLI) task, and one sentiment analysis task. The results showed that GreekBART outperforms the other models in both the classification and NLI tasks, demonstrating its superior ability to distinguish between different language classes and infer relationships between sentences. However, Greek-BERT exhibited stronger performance in the sentiment analysis task, suggesting that BERT-based models may still have an edge in specific, nuanced language tasks that require a deep understanding of emotional content.

In the second evaluation, GreekBART was compared to mBART 25, mBART 50, and BART-random models on two generative tasks, specifically focused on summarization. The results indicate that GreekBART's performance is comparable to that of BART-LARGE models (mBART 25 and mBART 50), underscoring its effectiveness in generating coherent and accurate text summaries. This puts GreekBART on par with some of the most advanced multilingual models available, particularly in tasks that require transforming or summarizing large amounts of text.

1.4. HuggingFace – Model Hub[1]

The HuggingFace Model Hub (https://huggingface.co/models) serves as a vast repository for LLMs, offering users the ability to both share and access a wide range of open-source models. As of now, the platform hosts over 1,000,000 open-source LLMs, including 921 models that specifically support the Greek language. These Greek-compatible models are either fine-tuned versions of existing LLMs or models that were originally trained on multilingual datasets, which include Greek as part of their language repertoire.

Text Classification	Token Classification	Table Question Answering	Zero-shot Classification	Translation	Summarization
51	46	0	21	**134**	6
Question Answering	**Feature Extraction**	**Text Generation**	**Text2Text Generation**	**Fill-Mask**	**Sentence Similarity**
2	33	**134**	52	62	62

Table 1. Existing opensource Greek-compatible LLMs included in HuggingFace

Table contains the amount of HuggingFace's Model Hub models (to date) that support Greek language in several Natural Language Processing tasks. As is apparent, a greater number of models support Translation and Text Generation task among the range of NLP tasks.

The HuggingFace Model Hub provides a diverse selection of models that support numerous languages and can be applied to a broad spectrum of Natural Language Processing (NLP) tasks. These tasks include but are not limited to text classification, text generation, question answering, translation, and summarization. Each model is designed to handle specific language tasks, enabling researchers and developers to select models tailored to their unique requirements.

Beyond NLP, the HuggingFace Model Hub also contains models for tasks that go beyond text-based applications. This includes models dedicated to Computer Vision, allowing users to engage with tasks such as image recognition, object detection, and image generation.

[1] https://huggingface.co/models

2. Greek Learner Corpus II (Tantos et al., 2023)

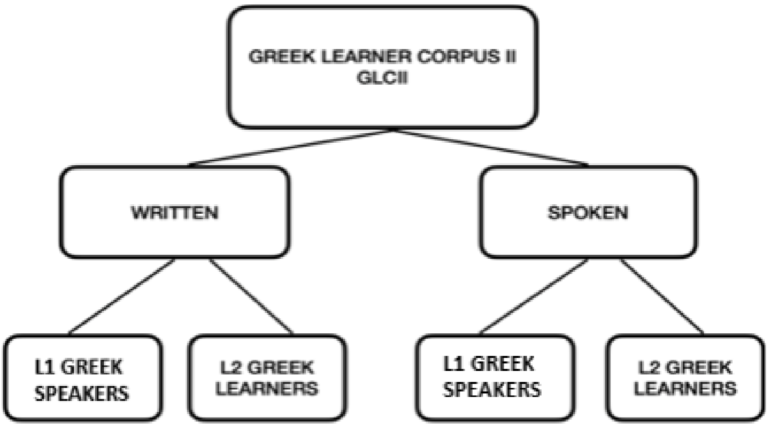

Figure 4 GLCII structural design (source: Tantos et al, 2023)

GLCII is the largest widely available learner corpus of Greek as a second language (L2). It is a growing learner corpus and comprised of written and spoken productions of adult L2 Greek learners, accompanied by a small control subcorpus with productions from native speakers of Greek. The corpus also represents a wide spectrum of proficiency levels, from beginner to advanced. The L2 Greek learners participating in GLCII attended Greek language courses in Greece or abroad and most of their productions come from instructed language learning context. The texts have been error-annotated for five fundamental grammatical categories: Agreement, Voice, Gender, Case and Aspect. Each category is equipped with a specific tagset to identify relevant error cases. Furthermore, GLCII includes extensive descriptive metadata, relevant to learner's linguistic profile (L1, proficiency level etc.), sociocultural profile, demographic context (sex, age, country of origin and educational level) and text and task related variables.

Proficiency level	**Written**		**Spoken**		**Total**	
	Texts	Word Tokens	Texts	Word Tokens	Texts	Word Tokens
A1	18	1997	-	-	18	1997
A2	125	16,139	42	~20,605	167	~36,744
B1	283	45,009	105	~51,510	388	~96,519
B2	480	90,548	86	~42,190	566	~132,738
C1	178	40,897	77	~37,775	255	~78,672

C2	17	5,120	8	~3,925	25	~9045
Total	**1101**	**197,713**	**318**	**~156,005**	**1419**	**~422,360**

Figure 5 GLCII (currently) data (source: Tantos et al., 2023)

GLCII, being the most comprehensive database of L2 Greek productions to date will serve as the basis dataset for fine-tuning the two foundation LLMs, Meltemi and GreekBART.

3. First steps for fine-tuning process

This section outlines the preliminary steps in fine-tuning the pretrained Greek-BERT and Meltemi models for the GED task, specifically focusing on L2 Greek. As mentioned in Section 1, fine-tuning is a supervised learning process that requires a labeled dataset tailored to the specific downstream task. However, since there is no existing dataset dedicated to GED for Greek, our first task was to create one.

To build this dataset, we utilized the GLCII, an open-access resource that provides a substantial amount of authentic language data relevant to the target linguistic variety. The GLCII already contains error-annotated texts, making it a valuable starting point for constructing a dataset for GED. However, the existing annotations required refinement to suit our specific needs.

For the already error-annotated texts, we conducted an exhaustive review to ensure the accuracy and consistency of the error labels. This process involved filtering out any erroneous or inconsistent annotations, adding new annotations for previously undetected errors, and consolidating duplicate tags.

Throughout this annotation process, we adhered closely to the established GLCII annotation scheme but made necessary modifications to enhance the dataset for our fine-tuning purposes. In addition to span annotations, we introduced a relational annotation layer specifically designed to capture Agreement errors more explicitly. Agreement errors, such as those involving subject-verb agreement or noun-adjective agreement, are now defined as relational errors between words. This approach allows for a more detailed and comprehensive representation of these errors, improving the quality of the training data. The tagset for these relational annotations aligns with the GLCII's tagset, but with added clarity to define such errors that express their relational nature, which will aid in the model's ability to detect them during training.

1. <<...δεν θα σώσουν **το** κόσμο...>> [=...will not save the world...]
2. <<...η μετανάστευση βοηθάει να **γεφυρώνει**...>> [=...migration helps to bridge...]

Annotating errors in L2 Greek poses unique challenges due to the complexity of Greek morphology. One significant issue is ambiguity in certain error cases, where multiple plausible annotations exist for a single error. The noun phrase *το κόσμο* in (1) presents two possible annotation options. In the first option, there is a gender disagreement between the article and the noun, since the article *το* has a neuter gender, while the following noun, *κόσμο* [=*world*], has been assigned a masculine gender by the learner. However, in the second interpretation, *κόσμο* could be interpreted as having been assigned a neuter gender and, therefore, there is not gender disagreement but a gender assignment case. Due to syncretism—the overlap between the accusative forms of masculine nouns ending in *-ος* and neuter nouns ending in *-ο* creates such ambiguity instances.

In addition to ambiguity, there are cases where multiple annotations are required at the same time. In (2), two annotation tags, one for Aspect and one for Voice, are necessary to capture the fact that the verb form *γεφυρώνει* [=*bridges*] is incorrectly marked with imperfective aspect and active voice, instead of its expected form.

For the error annotation process, we used INCEpTION (Klie et al., 2018), a widely recognized and publicly available annotation platform that allows relational span annotation. INCEpTION's flexible interface enabled the smooth integration of the relational annotation layer into our workflow, facilitating efficient error marking and management, and ensuring the creation of a structured, high-quality dataset.

The next stage involves dividing the annotated dataset into three subsets: training, evaluation, and test sets. This step is critical for properly assessing the performance of the fine-tuned models on the GED task, allowing us to measure their effectiveness in detecting grammatical errors in Greek learner productions. By structuring the data in this way, we aim to crate the bases for comparing the robustness of LLMs that are asked to detect and correct a wide range of grammatical errors specific to Greek as a second language.

The next steps in the fine-tuning process involve dataset tokenization, where the raw data is transformed into a format that the model can interpret. Specifically, the data is first split into tokens, then converted into numerical representations, and finally into tensors. For this, we will use the AutoTokenizer class from the Hugging Face Transformers library, which automatically selects the appropriate tokenizer based on the model architecture. Following this, we will utilize the Trainer class from the same library to train our model.

Key tasks include loading the Greek-BERT and Meltemi models and creating a TrainingArguments class instance that defines the required hyperparameters, such as the optimizer and learning rate. Last step is to set up an evaluation function and initialize a Trainer object, passing the model, training arguments, dataset, and evaluation function as input parameters.

4. Conclusion

Large Language Models (LLMs) have seen remarkable advancements in recent years. The development of Greek-specific models, such as GreekBART in 2023 and Meltemi in 2024, underscores both the growth of LLMs and the importance of language-specific models. However, many languages and language varieties remain significantly underrepresented in the pretraining datasets of widely used LLMs. As a result, these models are less effective at processing these less represented languages and language varieties compared to more dominant languages, like English. One such underrepresented variety is the output produced by second language learners.

To mitigate this issue, techniques like fine-tuning offer a practical solution by adapting LLMs to specific language varieties and tasks. Fine-tuning allows a pretrained model to acquire specialized *knowledge* about the variety it is trained on, making it better suited to handle the nuances of that variety. In our case, fine-tuning enables the model to internalize and process the specific features of second language (L2) learner output. This approach is more resource-efficient compared to training a model from scratch, as it leverages the knowledge gained during pretraining while focusing on adapting the model to a new task or language variety. For this process, high-quality datasets like the GLCII are invaluable, as they provide authentic and diverse L2 learner data essential for effective fine-tuning.

Fine-tuned LLMs designed for second language learning present numerous benefits for both students and educators. For students, a fine-tuned model for GED can offer immediate, accurate feedback tailored to their specific learning needs. Since the model is trained on data from other L2 Greek learners with similar language profiles, it can address the types of errors and challenges that students typically face, improving the relevance and usefulness of the feedback. Additionally, using instruct versions of such models can simulate real-world communication scenarios outside the classroom, allowing students to practice and improve their language skills during self-study, thus maximizing the effectiveness of their study time at home.

For educators, fine-tuned LLMs can serve as powerful tools for monitoring student progress. These models can provide detailed insights into learners' performance, identifying common errors and highlighting areas where individual students may need more targeted instruction. This enables teachers to develop more personalized and effective teaching materials, thus enhancing the overall learning experience. By reducing the time spent on repetitive grading tasks and offering valuable diagnostic feedback, fine-tuned LLMs not only improve learning outcomes but also help educators manage their workload more efficiently.

References

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. arXiv:1706.03762 [cs.CL]

Tantos A., N. Amvrazis & E. Drakonaki. (2023). Greek Learner Corpus II (GLCII): Design and development of an online corpus for L2 Greek. *Journal of Applied Linguistics, 36*, 125– 151. doi:10.26262/jal.v0i36.9915

Schulhoff, S., Ilie, M., Balepur, N., Kahadze, K., Liu, A., Si, C., Li, Y., Gupta, A., Han, H., Schulhoff, S., Dulepet, P., Vidyadhara, S., Ki, D., Agrawal, S., Pham, C., Kroiz, G., Li, F., Tao, H., Srivastava, … Resnik, P. (2024). The prompt report: A systematic survey of prompting techniques. arXiv:2406.06608 [cs.CL] CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Amatriain, X. (2024). Prompt design and engineering: Introduction and advanced methods. arXiv preprint arXiv:2401.14423 [cs.SE] CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Koutsikakis, J., Chalkidis, I., Malakasiotis, P., & Androutsopoulos, I. (2020). Greek-bert: The greeks visiting sesame street. In C. Spyropoulos, I. Varlamis, I. Androutsopoulos & P. Malakasiotis (Eds.), *11th Hellenic Conference on Artificial Intelligence (SETN 2020)*, (pp. 110-117). Association for Computing Machinery. doi:10.1145/3411408.3411440

Voukoutis, L., Roussis, D., Paraskevopoulos, G., Sofianopoulos, S., Prokopidis, P., Papavasileiou, V., Katsamanis, A., Piperidis, S., Katsouros, V. (2024). Meltemi: The first open Large Language Model for Greek. arXiv:2407.20743 [cs.CL] CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Evdaimon, I., Abdine, H., Xypolopoulos, C., Outsios, S., Vazirgiannis, M., Stamou, G. (2023). GreekBART: The First Pretrained Greek Sequence-to-Sequence Model. arXiv:2304.00869 [cs.CL] CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Klie, J.-C., Bugert, M., Boullosa, B., Eckart de Castilho, R. & Gurevych, I. (2018). The INCEpTION Platform: Machine-Assisted and Knowledge-Oriented Interactive Annotation. *Proceedings of System Demonstrations of the 27th International Conference on Computational Linguistics (COLING 2018)*

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.D.L., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.-A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T. & Sayed, W.E. (2023). Mistral 7B. arXiv preprint arXiv:2310.06825 [cs.CL] CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Koehn, P. (2005). Europarl: A Parallel Corpus for Statistical Machine Translation. *Proceedings of Machine Translation Summit X: Papers,* 79–86.

Outsios, S., Skianis, K., Meladianos, P., Xypolopoulos, C. & Vazirgiannis, M. (2018). Word embeddings from large-scale greek web content. arXiv preprint arXiv:1810.06694 [cs.CL]

Suárez, P.J.O., Sagot, B. & Romary, L. (2019). Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. 7th Workshop on the Challenges in the Management of Large Corpora (CMLC-7), Cardiff, United Kingdom. 10.14618/IDS-PUB-9021. fhal-02148693 CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Abadji, J., Suarez, P.O., Romary, L. & Sagot, B. (2022). Towards a Cleaner Document-Oriented Multilingual Crawled Corpus. *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 4344–4355). European Language Resources Association.

Liu, T. & Low, B. K. H. (2023). Goat: Fine-tuned llama outperforms gpt-4 on arithmetic tasks. arXiv preprint arXiv:2305.14201 [cs.LG] CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/)

Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M. & Le, Q.V. (2022). Finetuned language models are zero-shot learners. Tenth International Conference on Learning Representations (ICLR 2022)