CLiC-it 2025

The Eleventh Italian Conference on Computational Linguistics

Proceedings of the Conference

September 24-26, 2025

These proceedings have been published in the CEUR Workshop Proceedings series.
The original papers are available at: https://ceur-ws.org/Vol-4112.
The papers are mirrored in the ACL Anthology.

ISBN 979-12-243-0587-3

Table of Contents

Preface to the CLiC-it 2025 Proceedings

Cristina Bosco[1], Elisabetta Jezek[2], Marco Polignano[3] and Manuela Sanguinetti[4]

[1]University of Torino
[2]University of Pavia
[3]University of Bari "Aldo Moro"
[4]University of Cagliari

The Italian Conference on Computational Linguistics (CLiC-it) is the yearly conference organized by the *Associazione Italiana di Linguistica Computazionale* (Italian Association of Computational Linguistics, AILC). Its main goal is to promote and spread original and high-quality research on the diverse aspects concerning the automatic processing of natural language, both spoken and written. The eleventh edition of CLiC-it took place in Cagliari, from the 24th to the 26th of September 2025.

In line with previous editions, submissions to the conference could be of two types: regular papers, featuring original and unpublished contributions, and non-archival research communications, consisting of papers accepted in 2024 and 2025 by major publication venues, namely the major international Computational Linguistics (CL) conferences (workshops excluded) or international journals. Regular paper submissions were assigned to thirteen Senior Program Chairs based on the general area that covered the paper's topic. Paper assignments to reviewers were managed individually by the single Senior PCs, though resorting to a global pool of 151 available reviewers. We have received 138 submissions of regular papers, hitting once more the record number of submissions even compared to the previous edition in 2024, where regular papers submitted were 133. This confirms the vitality and growth of the Italian Computational Linguistics community. Along with regular papers, we also received 22 research communications. Among the regular submissions, 113 were accepted for presentation at the conference, resulting in a 81.8% acceptance rate, with respect to the 85.7% rate of CLiC-it 2024. Out of these, 55 were accepted as oral presentations and 58 as posters. After the author notification was sent, 4 papers were withdrawn by the authors themselves. As a result, the conference featured a total of 55 oral presentations and 54 posters. Finally, of the 22 research communications submitted – a clear sign of the vitality and quality of the research carried out within the community – 14 were included for poster presentation in a dedicated session. The selection was not based on an additional review process, but rather on the venue of publication. Even in this case, two research communications were withdrawn after the notification, hence 12 posters were presented at the conference.

The program also included two keynote talks, by **Karen Fort** (University of Lorraine) and **Edoardo Maria Ponti** (University of Edimburgh/NVIDIA):

- Karen Fort gave a talk titled "Large Language Models: the challenge of evaluation": *In the past five years or so, Natural Language Processing has witnessed a revolution. Not only have Large Language Models (LLM) submerged the domain, but they also invaded our societies: our systems now have real users and an impact on their lives. This dramatic change happened so fast that we -the research community- are still trying to catch up, especially concerning the evaluation of the real capabilities of these tools. In this presentation, I'll show what are the flaws of the present LLMs' evaluation and how ethics is a powerful leverage to improve it.*
- the talk by Edoardo Maria Ponti was titled "A Blueprint for Foundation Models with Adaptive Tokenization and Memory": *Foundation models (FMs) process information as a sequence of internal representations; however, the length of this sequence is fixed and entirely determined by tokenization. This essentially decouples representation granularity from information content, which exacerbates the deployment costs of FMs and narrows their "horizons" over long sequences. What if, instead, we could free FMs from tokenizers by modelling bytes directly, while making them faster than current tokenizer-bound FMs? To achieve this goal, I will show how to: 1) learn tokenization end-to-end, by dynamically pooling representations in internal layers and progressively learning abstractions from raw data; 2) compress the KV cache (memory) of Transformers adaptively during generation without loss of performance; 3) predict multiple bytes per time step in an efficient yet expressive way; 4) retrofit existing tokenizer-bound FMs into byte-level FMs through cross-tokenizer distillation.*

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
cristina.bosco@unito.it (C. Bosco); elisabetta.jezek@unipv.it (E. Jezek); marco.polignano@uniba.it (M. Polignano); manuela.sanguinetti@unica.it (M. Sanguinetti)

By blending these ingredients, we may soon witness the emergence of new, more efficient architectures for foundation models.

One session of the conference was devoted to a discussion with **Paola Merlo**, conceived as a space for critical discussion on key areas of research in the fields of Computational Linguistics and Natural Language Processing, with a focus on both theoretical developments and applied scenarios within these disciplines. The discussion took place in an interview format. With the aim of promoting active and inclusive participation, a call for interest was launched, open to all interested participants - with a special focus on young researchers - to collect questions in advance to be addressed to Paola Merlo during the interview.

The conference was also preceded by a tutorial held by **Sandro Pezzelle** (University of Amsterdam) titled "*Language-and-vision models: From image-language alignment to storytelling and narration*". The tutorial aimed at providing an accessible yet in-depth overview of language-and-vision models, ranging from traditional modular pipelines to the latest end-to-end pre-trained systems (VLMs). The tutorial introduced foundational concepts and architectures, then focusing on recent approaches and evaluation challenges.

AILC also renewed its support to the **Emanuele Pianta Award** for the Best Master's Thesis defended at any Italian university between August 1st 2024 and July 31st 2025, and addressing a topic in computational linguistics or its applications. This year we received 9 candidate theses for the award. Of these, 2 were not further considered for evaluation due to incomplete submission. The candidate theses were evaluated by a jury composed of a current chair of CLiC-it, specifically Elisabetta Jezek was designated for this role, a co-chair of the past edition, i.e. Rachele Sprugnoli (University "Cattolica del Sacro Cuore" of Milan), and a further member of the AILC Board, i.e. Danilo Croce (University "Tor Vergata" of Rome). The winner was awarded by the members of the jury during the closing session of the conference.

We would like to thank all the **institutions** involved in the organization of the conference and the people of these institutions that worked with us for creating a successful event. For the logistic support, our thanks go to the Faculty of Economics, Law and Political Sciences of the University of Cagliari, that hosted the conference in the Sant'Ignazio Campus, and Valentina Deidda in particular, who kindly assisted us in all the technical and logistic aspects concerning the organization of the event. For the organizational and financial support, our thanks also go to the Department of Mathematics and Computer Science of the University of Cagliari, whose researchers were involved in the development and management of the website and whose students worked hard to help run things smoothly during the conference.

Our gratitude also goes to our **corporate supporters** for their generous provision of the financial resources and services that made this event possible: Aptus.AI, Talia, Aequa-tech, Almawave, Domyn, Elra, Logogramma, Prometeia, and Translated.

We express our deepest gratitude also to all Senior Program Chairs, all members of the Program Committee, and all participants, who contributed to the success of the event. Chairs and reviewers are named in the following pages.

Our final and special thanks go to **AILC**'s Board, whose members constantly supported us, giving guidance and invaluable assistance throughout the whole organization of the event, and to the whole AILC's association members that make this scientific event more interesting and richer every year.

Cagliari, September 2025

Conference Chairs

- **Cristina Bosco**, University of Torino
- **Elisabetta Jezek**, University of Pavia
- **Marco Polignano**, University of Bari "Aldo Moro"
- **Manuela Sanguinetti**, University of Cagliari

Local Committee

- **Maurizio Atzori**, University of Cagliari
- **Andrea Loddo**, University of Cagliari
- **Davide Antonio Mura**, University of Cagliari
- **Alessandro Pani**, University of Cagliari
- **Alessandra Perniciano**, University of Cagliari
- **Luca Zedda**, University of Cagliari

Proceeding Chairs

- **Francesca Grasso**, University of Torino
- **Andrea Zaninello**, Fondazione Bruno kessler

Publicity and Data Chairs

- **Alessandro Bondielli**, University of Pisa
- **Mirko Lai**, University of Eastern Piedmont

Booklet

- **Alessandra Perniciano**, University of Cagliari

Webmasters

- **Maurizio Atzori**, University of Cagliari
- **Andrea Loddo**, University of Cagliari

Registration Management

- **Sara Barcena**, freelance designer
- **Manuela Speranza**, Fondazione Bruno Kessler

Senior Program Chairs

- **Elisa Bassignana**, IT University of Copenhagen
- **Pierluigi Cassotti**, University of Gothenburg
- **Simone Conia**, University of Rome "La Sapienza"
- **Elisa Di Nuovo**, Joint Research Centre European Commission, Ispra
- **Claudiu Daniel Hromei**, University of Rome "Tor Vergata"
- **Antonio Origlia**, University of Naples "Federico II"
- **Ludovica Pannitto**, University of Bologna
- **Beatrice Savoldi**, Fondazione Bruno Kessler
- **Gabriele Sarti**, University of Groningen
- **Lucia Siciliani**, University of Bari
- **Irene Siragusa**, University of Palermo
- **Rossella Varvara**, University of Torino/University of Pavia
- **Alessandro Vietti**, University of Bolzano

Program Committee

Oscar Araque, Serena Auriemma, Pier Balestrucci, Matilde Barbini, Alberto Barrón-Cedeño, Pierpaolo Basile, Valerio Basile, Monica Berti, Siddharth Bhargava, Arianna Bienati, Andrea Bolioli, Alessandro Bondielli, Giovanni Bonetta, Tommaso Bonomo, Federico Borazio, Federico Boschetti, Sofia Brenna, Luca Brigada Villa, Dominique Brunato, Davide Buscaldi, Elena Cabrio, Sara Candussio, Luca Capone, Franco Alberto Cardillo, Alessio Cascione, Silvia Casola, Giuseppe Castellucci, Camilla Casula, Flavio Massimiliano Cecchini, Giuseppe G.A. Celano, Mauro Cettolo, Emmanuele Chersoni, Cristiano Chesi, Francesca Chiusaroli, Alessandra Teresa Cignarella, Lorenzo Cima, Davide Colla, Gloria Comandini, Claudia Roberta Combei, Lina Conti, Serena Coschignano, Danilo Croce, Francesco Cutugno, Lorenzo De Matteo, Angelo Maria Del Grosso, Pietro Dell'Oglio, Maria Pia Di Buono, Andrea Di Fabio, Maria Di Maro, Luca Dini, Luca Ducceschi, Andrea Esuli, Alfio Ferrara, Elisabetta Fersini, Komal Florio, Simona Frenda, Francesca Frontini, Achille Fusco, Eleonora Ghizzota, Aivars Glaznieks, Claudiu Hromei, Alina Karakanta, Fahad Khan, Tiziano Labruna, Mirko Lai, Alberto Lavelli, Gianluca Lebani, Els Lefever, Alessandro Lenci, Elisa Leonardelli, Soda Marem Lo, Agnese Lombardi, Marco Madeddu, Bernardo Magnini, Francesco Mambrini, Chiara Manna, Raffaele Manna, Marta Marchiori Manerba, ndrea Marra, Claudia Marzi, Arianna Masciolini, Alessandro Mazzei, Enrico Mensa, Alessio Miaschi, Simonetta Montemagni, Johanna Monti, Luca Moroni, Elio Musacchio, Benedetta Muscato, Vivi Nastase, Roberto Navigli, Nicole Novielli, Debora Nozza, Antonio Origlia, Teresa Paccosi, Francesca Padovani, Alessio Palmero Aprosio, Sara Papi, Lucia Passaro, Marco Passarotti, Viviana Patti, Matteo Pellegrini, Nicolò Penzo, Federico Pianzola, Vito Pirrelli, Roberto Pirrone, Massimo Poesio, Mattia Proietti, Valeria Quochi, Giulia Rambelli, Federico Ranaldi, Leonardo Ranaldi, Irene Russo, Daniel Russo, Andrea Santilli, Loredana Schettino, Giulia Speranza, Rachele Sprugnoli, Marco Antonio Stranisci, Carlo Strapparava, Alice Suozzi, Luigi Talamo, Fabio Tamburini, Benedetta Tessa, Davide Testa, Sara Tonelli, Giulia Venturi, Guido Vetere, Serena Villata, Vincenzo Norman Vitale, Roberto Zamparelli, Andrea Zaninello, Roberto Zanoli.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly for grammar and spelling check. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

Towards the Semi-Automated Population of the Ancient Greek WordNet

Beatrice Marchesi[1,*], Annachiara Clementelli[1], Andrea Maurizio Mammarella[1], Silvia Zampetta[1], Erica Biagetti[1], Luca Brigada Villa[1], Virginia Mastellari[1], Riccardo Ginevra[2], Claudia Roberta Combei[3] and Chiara Zanchi[1]

[1]*Università degli Studi di Pavia*
[2]*Università Cattolica del Sacro Cuore*
[3]*Università degli Studi di Roma "Tor Vergata"*

Abstract

This paper explores the employment of LLMs, specifically of Mistral-Nemo, in the semi-automatic population of the Ancient Greek WordNet synsets. Several approaches are investigated: zero-shot, few-shots, and fine-tuning. The results are compared against an English baseline. Zero-shot approach yields the highest accuracy, while fine-tuning leads to the highest number of potential synonyms. Our analysis also reveals that polysemy and PoS play a role in the model's performance, as the highest scores are registered for polysemous words and for verbs and nouns. The results are encouraging for the application of such approaches in a human-in-the-loop scenario, since human validation still proves crucial in ensuring the accuracy of the results.

Keywords

Lexical semantics, synonym generation, LLMs, Ancient Greek, WordNet

1. Introduction

In this paper, we explore the application of Large Language Models (LLMs) for populating the synsets of the Ancient Greek WordNet (AGWN) and assessing the extent to which these models can support such a task.

WordNet is a lexical resource that organizes word meanings by groups of quasi-synonymous words connected to each other in a network structure ([1]). The first WordNet was developed for English at Princeton University by George Miller and Christiane Fellbaum ([2], [3], [4]). Originally developed within a project in psycholinguistics, it gradually evolved into a tool for computational lexical semantics. The development of such semantic networks was subsequently extended to languages beyond English, beginning with modern languages (e.g., [5]) and later including ancient ones as well, such as Latin, Ancient Greek, Sanskrit and Old English ([6], [7], [8], [9], [10]).

The building blocks of WordNets are synsets, that is, groups of cognitive synonyms, each associated with a short definition and an ID-number ([1]). WordNets are designed to represent both synonymy and polysemy, via assignment to the same synset or to multiple synsets, respectively. For example, the Ancient Greek nouns *apaugasmós, aíglē, kataúgasma, phōtḗr, apaúgasma, periphéggeia, augasmós, bolḗ, kiéllē*[1] all belong to the synset n#03874115 'the quality of being bright and sending out rays of light', indicating that they are at least partially synonyms[2]. In addition, lemmas can be assigned to multiple synsets, which indicates polysemy. This is the case for *aíglē*, which also appears in the synsets n#03874461 'an appearance of reflected light' and n#03690420 'brilliant radiant beauty: "the glory of the sunrise"'. Furthermore, synsets are connected via semantic relations such as hyponymy, hyperonymy, and meronymy, whereas lexemes are related to one another via lexical relations, primarily derivation.

Drawing from a previous collaboration with the University of Pavia ([13]), the first version of the AGWN was developed in 2014 as the result of an international collaboration between the Institute of Computational Linguistics "Antonio Zampolli" (Pisa), the Perseus Project, the Open Philology Project, and the Alpheios Project. It

CLiC-it 2025: Eleventh Italian Conference on Computational Linguistics, September 24 — 26, 2025, Cagliari, Italy
*Corresponding author.
beatrice.marchesi03@universitadipavia.it (B. Marchesi); annachiara.clementelli01@universitadipavia.it (A. Clementelli); andreamaurizio.mammarella01@universitadipavia.it (A. M. Mammarella); silvia.zampetta01@universitadipavia.it (S. Zampetta); erica.biagetti@unipv.it (E. Biagetti); luca.brigadavilla@unipv.it (L. Brigada Villa); virginia.mastellari@unipv.it (V. Mastellari); riccardo.ginevra@unicatt.it (R. Ginevra); claudia.roberta.combei@uniroma2.it (C. R. Combei); chiara.zanchi@unipv.it (C. Zanchi)

[1]Note that in the experiment both the inputs and the outputs of the model were written in the Greek alphabet. In this paper, however, all Ancient Greek lemmas are transliterated and provided with translations supplied by the LSJ lexicon [11].
[2]Synsets do not group together only 'absolute synonyms', i.e., words that are interchangeable in all possible contexts, but also words that are similar in meaning limited to certain contexts ([2]: 241, [12].)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2025-11-30

was initially constructed using digitized Greek-English lexica from the Perseus Project, linking the Greek word of each extracted bilingual pair to every synset in the Princeton WordNet ([3]) in which the English member of the pair appeared. This method, known as the *expand method* ([5]), has been commonly adopted in the development of several modern WordNets ([14]), largely due to the extensive richness and detail of the Princeton WordNet. However, it presents challenges typical of using English as a pivot language, as well as difficulties specific to mapping concepts across culturally and historically distant traditions. In the case of the AGWN, synsets were also aligned with the Italian section of the MultiWordNet ([15]), ItalWordNet ([16]), and with the Latin WordNet ([6]). A subset of synsets was used to evaluate the automatic extraction process and erroneous alignments were removed by filtering out anachronistic domains. This version of the AGWN included approximately 35,000 lemmas—roughly 28% of the estimated 120,000 lemmas in the entire Ancient Greek lexicon. Coverage was significantly higher for the Homeric lexicon (69%), owing to the incorporation of Autenrieth's *Homeric Dictionary* in the construction of the resource (see [7] for details).

The work on the AGWN continues in the framework of the PRIN project *Linked WordNets for Ancient Indo-European Languages*, whose aim is to harmonize three WordNets for Ancient Greek, Latin, and Sanskrit, and expand their coverage in terms of the number of annotated words and populated synsets ([9], [17]).

While various methods have been proposed for the automatic population of synsets, their outputs typically still require substantial manual validation. For instance, word embeddings have been employed to identify lexical relations absent from existing WordNets for Ancient Greek ([18]), Sanskrit ([19]), and Latin ([20]; see [21] for an overview). Given that fully manual synset population is highly time-consuming, a further aim was later added to the project *Linked WordNets for Ancient Indo-European Languages*: the training and testing of LLMs for the automatic population of synsets of ancient languages. These models are intended to be integrated into the current annotation platform to suggest potential synonyms to annotators, who will then manually validate the LLM generations.

The first experiment with LLMs, conducted on Latin ([21]), aimed to compare zero-shot, few-shot, and fine-tuning approaches against an English baseline. Quantitative analysis showed marked improvements from zero-shot to fine-tuning approaches, with the latter outperforming the English baseline. Qualitative evaluation revealed stronger performance with verbs and with lemmas belonging to relatively well-populated synsets. While the results were encouraging, they highlighted the need for better performance across various parts of speech and degrees of polysemy. These goals are pursued in the present paper, which extends the experiment to Ancient Greek.

The paper is organized as follows. In Section 2 we describe our data and methodology, discussing the creation of the dataset (2.1), the zero-shot approach (2.2), the few-shot approach (2.3), and the fine-tuning processes performed using the LoRA technique (2.4). In Section 3 we report the results of the experiment, which are discussed from both a quantitative (3.1) and a qualitative (3.2) perspective. Section 4 concludes the paper.

2. Data and Methodologies

The experiment [3] followed three distinct methodological phases, namely zero-shot prompting, few-shot prompting, and fine-tuning. This progression was introduced to evaluate the effectiveness of different approaches for the given task and determine the advantages and disadvantages of each strategy.

Furthermore, an English baseline was established to validate the results of this study, in order to explore the model's responsiveness to this specific task and to examine how cross-linguistic differences might influence its performance.

The pretrained model used in all stages of the experiment is Mistral-NeMo[4], a multilingual open source model selected because of its balance between performance and efficiency, which results optimal for fine-tuning.

2.1. Datasets

The testing data used in the experiment consists of two datasets, one made up of (chiefly) monosemous lemmas and the other of polysemous lemmas. This distinction follows the work of [21], in which the distinction of the two datasets was based on the number of lemmas associated to the synsets: the so-called polysemous dataset was formed by well-populated synsets, each containing 15 mainly polysemous lemmas, while the so-called monosemous dataset was made up by less populated synsets containing at least two monosemous lemmas. However, in this work the datasets were manually crafted, since the annotated data in the AGWN are too scarce to allow for the same approach: lemmas possessing just one meaning according to the LSJ lexicon ([11]) were collected in the monosemous dataset, while lemmas associated to multiple meanings constitute the polysemous dataset. Each of the datasets is composed of 40 lemmas, equally divided

[3] The datasets, code, and data used for this experiment are provided in a repository at https://github.com/unipv-larl/llms-ag.
[4] https://mistral.ai/news/mistral-nemo, https://huggingface.co/mistralai/Mistral-Nemo-Instruct-2407

among the four PoS types included in WordNets (10 verbs, 10 nouns, 10 adjectives, and 10 adverbs).

To validate the results against a benchmark, an English baseline (EB) dataset was created. Considering that the English baseline serves as a benchmark to highlight differences in performance between a high-resource modern language such as English and Ancient Greek, a substantial gap between the results for the two target languages is to be expected. The English baseline dataset maintains the distinction between "monosemous" and "polysemous" sets, and its characteristics are the same as those of the test dataset. Thus, the included lemmas have roughly the same meanings as the Ancient Greek words, since they consist of translations and are balanced for PoS. During the translation of the Ancient Greek dataset into English, particular care was taken to preserve the distinctions between the datasets. Lemmas from the monosemy dataset were translated using roughly monosemous English words, while those from the polysemy dataset were rendered with mainly polysemous equivalents.

The fine-tuning dataset was created by extracting data from back-translation dictionaries, based on the assumption that such dictionaries provide, for any given entry in a modern language, a list of Ancient Greek words that can be used in context to translate that entry, that is, contextual synonyms. An example of a back-translation dictionary entry is offered below:

- **Accusation** (subs.): P. *katēgoría, hē*, *katēgórēma, tó*, P. and V. *aitía, hē*, *aitíama, tó*, *énklēma, tó*, V. *epíklēma, tó* ([22]).

Through a series of processing and cleaning operations, a dictionary of Ancient Greek synonym sets was extracted from the English-Greek Dictionary ([22]) and the Deutsch-Griechisches Wörterbuch ([23]), merging the results obtained from each dictionary to avoid overlap. It is important to note that the digital versions of these back-translation dictionaries were obtained through OCR (Optical Character Recognition), which - while generally accurate for modern languages written in the Latin script - yields sub-optimal results for Ancient Greek, often producing incorrectly digitized data and, consequently, inexact outputs. To address this problem, a series of cleaning operations was performed, from encoding normalization to checking the lemmas against the entries of the Brill Dictionary ([24]) to exclude incorrect or non-existent words. Such cleaning procedures ensure that the assembled dictionary only contains existing Ancient Greek words in their lemmatized form and that each set of synonyms exclusively features lemmas pertaining to the same PoS. An example of the synonym sets resulting from the data collection procedure is presented below:

- ***phrikṓdēs*** (awe-inspiring): *ouránios* (heavenly), *theîos* (divine), *deinós* (wondrous).

The resulting dataset in JSONL format was made up of 5,458 sets of synonyms with a mean number of 16 synonyms each (minimum 1, maximum 315 for the lemma *peribállō* (throw around)), thus divided across PoS: 2946 nouns (54%), 1372 verbs (25%), 955 adjectives (18%) and 185 adverbs (3%)[5].

The aim of the experiment with Latin WordNet ([21]) was to explore the outcomes and benefits of automating WordNet annotation by fine-tuning a model with data extracted from the WordNet itself. The assumption was that training a model on data of the same type and with the same structure of the desired output might lead to improved results, creating a virtuous feedback loop in which WordNet data are directly used to generate new data for WordNet population. Although AGWN does not contain sufficient annotated data to provide a suitable training dataset and to support the exact same approach as [21], this work is based on the same assumption, since the data that was collected for fine-tuning shares the same structure and properties of the data in the WordNet, as previously discussed.

2.2. Zero-Shot Approach

The first approach of the experiment is zero-shot (ZS) learning. This strategy tests the generalization potential and performance of models in tasks for which they were not specifically trained, since "no demonstrations are allowed, and the model is only given a natural language instruction describing the task" ([25]: 7). Indeed, models pre-trained on various and general datasets are usually able to generalize across new tasks, thus saving resources needed to create labeled data for additional training or demonstrations ([26]).

Compared to other approaches, zero-shot learning presents several drawbacks, including difficulty with complex tasks and lower accuracy, as outputs may lack precision or contextual relevance. Moreover, it is highly sensitive to prompt framing, which plays a crucial role in this setting ([27]).

As the first stage of the experiment, the zero-shot strategy was applied for both the Ancient Greek dataset and the English baseline. The prompts were tailored to each language and followed the best practices of prompt engineering, such as assigning a persona, specifying the desired output format, and organizing assertions as a bullet list ([28]; [29]). For the complete prompts, see A.1 and A.2.

2.3. Few-Shot Approach

In the few-shot (FS) setting, some examples demonstrating the expected output, its format, and style are given

[5]The data collected for fine-tuning will be imported in the AGWN, to help with the automatic population of the resource.

to the model to enhance performance, helping it understand the reasoning required for the new task ([25]). This approach has been proven to generally outperform zero- and one-shot learning ([25]; [30]), especially in structured and complex tasks, such as synonym generation. Compared to fine-tuning, this method proves cost-effective because the weights of the model are left unchanged, sparing a computationally intensive process, and only a small set of labeled items is needed, which is convenient in cases of scarcity of data ([27]: 24). However, this strategy is strongly dependent on careful prompt engineering and on suitable and verified examples. Therefore, particular attention is needed when designing the prompts ([31]: 3). As for prompt engineering best practices, performance has been proven to increase the more similar the examples are to testing data. The choice of examples also seems to have a great effect on the output ([27]: 16).

To test this approach on the Ancient Greek dataset, an ad-hoc prompt was created by maintaining the basic structure of the zero-shot prompt and adding a set of eight examples featuring the same structure of the desired output. The examples are equally divided into roughly monosemous and polysemous word sets and are balanced for PoS, so that for each of the four PoS, two lemmas are provided, that is, one monosemous, the other one polysemous. The examples added to the few-shot prompt are listed in A.3.

2.4. Fine-Tuning with LoRA

A recent trend with demonstrated advantages is to adapt large-scale pre-trained language models to specific downstream tasks. Indeed, a first stage of generative pre-training leads to gaining a greater world and language knowledge and, consequently, to an improved performance. Then, the following fine-tuning (FT) on domain-specific labeled data updates the pre-trained parameters with a new training cycle to adapt the model to the task at hand. This combination of unsupervised pre-training and supervised fine-tuning results in a semi-supervised approach able to construct a universal representation, which can be applied to a wide array of tasks ([32]: 2).

Although fine-tuning greatly enhances model performance, it is very resource-intensive. Some strategies were explored to mitigate this issue, such as LoRA (Low-Rank Adaptation), which is a PEFT (Parameter-Efficient Fine-Tuning) method that makes fine-tuning more parameter- and compute-efficient by freezing the pre-trained model's parameters and adapting only a subset of weight matrices. This method proves to be highly efficient compared to traditional fine-tuning, especially with regard to memory and storage ([33]: 5), meeting and sometimes surpassing the baselines, without increases in inference times ([33]).

The final step of the experiment involved fine-tuning a task-specific model. This was achieved by fine-tuning the quantized Mistral-NeMo model, which was loaded in 8-bit format to optimize computational efficiency, using the previously described fine-tuning dataset on a GPU node of an HPC cluster. LoRA was used to optimize fine-tuning, setting the low-rank matrix dimension to 8 and the scale factor lora_alpha to 16, with a dropout of 10%. The dataset was split into training (80%) and validation (20%), and the training was set for five epochs with a learning rate of 1e-4. An early stopping mechanism with a patience of one epoch was established to avoid overfitting, and a parameter was set to save the model with the lowest value of validation loss, which corresponded to the output of the fourth epoch. The metrics calculated during fine-tuning over the five epochs of training are presented in Table 1.

Table 1
Fine-tuning metrics over the five epochs of training. For each metric, the best value is highlighted in bold type.

	1	2	3	4	5
Training loss	1,2943	1,4099	**1,1478**	1,2232	1,1855
Validation loss	1,4814	1,4366	1,4137	**1,4087**	1,4100
Training mean token accuracy	0,6587	0,6262	0,6597	0,6720	**0,7206**

The overall loss trend is descending, even if gradually, both in training and in validation, and the accuracy values are increasing. Overall, the metrics show that the training was conducted successfully and without overfitting.

3. Results and Discussion

The validation of the results took place in two steps. The first step was to automatically lemmatize each word using greCy ([34]), so that even inflected forms generated by the model are traced back to the corresponding lemma. Notably, this pre-processing step is pointless in the case of hallucinations or incorrect forms (for a more detailed discussion, see 3.2.1 and 3.2.2). It is worth pointing out that the lemmatization, while correct in most cases, was not always impeccable (e.g., *theoí* (gods, masculine nominative plural) > *theoí* (FS)).

After lemmatization, three human annotators[6] validated the results, determining for each generated item if it constituted a potential synonym of the input word. In

[6]The three annotators are all students of the MA program in Linguistics at the University of Pavia with a BA Degree in Classics.

cases of disagreement between the annotators, the matter was resolved through discussion until an agreement was reached. The inter-annotator agreement, measured with Fleiss' Kappa ([35]), reached a value of 0.71 on the Ancient Greek data and 0.66 on the English data, both of which fall under the label of good to substantial agreement. For the purposes of this work, the concept of synonymy is interpreted in a shallow and contextual sense, consistent with the framework upon which the WordNet architecture is based (see footnote 2). Thus, words whose meaning is similar enough that they might be assigned to the same synset are considered potential synonyms, as in 1.

1 ***anankázō***: rule, hold sway.
kratéō: force, compel.

The results are analyzed both from a quantitative and a qualitative perspective, and the analysis is carried out by comparing the different approaches employed, which are bench-marked against the English baseline. Regarding the quantitative data discussed in Section 3.1, the performance of each of the approaches is evaluated through the metrics of accuracy, similarity, number of generated outputs, and potential synonyms.

3.1. Quantitative Analysis

The results of the quantitative analysis are shown in Table 2, which displays the values of the metrics for each of the approaches, both providing the overall scores and distinguishing between the polysemous and the monosemous datasets.

Table 2
Metrics comparison (acc: accuracy, sim: similarity, n_gen: number of generated outputs, p_syn: number of potential synonyms). For each row, the best scores, excluding those of the EB, are highlighted in bold type to facilitate comparison across approaches for Ancient Greek synonym generation.

		acc	sim	n_gen	p_syn
Overall	EB	90%	.377	167	151
	ZS	**30%**	**.261**	116	34
	FS	5%	.099	169	9
	FT	11%	.077	**403**	**43**
Polysemy	EB	98%	.407	85	83
	ZS	**40%**	**.296**	63	24
	FS	7%	.066	61	4
	FT	13%	.113	**288**	**38**
Monosemy	EB	83%	.347	82	68
	ZS	**19%**	**.226**	53	**10**
	FS	5%	.132	108	5
	FT	4%	.041	**115**	5

As for the similarity metric, cosine similarity was computed using pre-trained Word2vec embeddings based on a skip-Gram model for both English[7] and Ancient Greek[8]. In a task such as synonym generation this metric is useful in determining if the output might be a valid synonym to the target word based on semantics and distribution. However, one limitation is represented by out-of-vocabulary (OOV) terms, meaning that in some cases, for both English and Ancient Greek, the metric fails to capture the actual similarity between the generated output and the input lemma, as one or both of the two words are not contained in the embedding dictionary, such as in 2.a and 2.b:

2.a **gourmand**: *epicure*. Similarity: 0.

2.b ***katasparássō*** (tear in pieces): *katagnúō* (break in pieces). Similarity: 0.

While the issue of OOVs affects both English and Ancient Greek, the latter is more severely impacted by this problem due to the more limited size of the embedding dictionary, thus the similarity values for Ancient Greek tend to be underestimated compared to the English baseline.

As shown in Table 2, the two datasets of the English baseline score the highest values in accuracy, similarity, total, and mean of potential synonyms. The results highlight that the model reaches a high performance in the task at hand, even in a zero-shot setting without task-specific demonstrations or training. This result indicates that the generalization potential of the model is quite high for a high-resource language such as English.

As for the zero-shot approach, the first step of the experiment shows a much lower performance compared to the English baseline, across all metrics. Considering that pre-trained models have much less data available for Ancient Greek compared to modern languages such as English, the drop in performance and in the number of generations is to be expected.

Considering now the few-shot approach, the results show an unexpected drop in performance compared to the zero-shot strategy. Indeed, the instructions given in the prompt apparently do not help the model, but rather affect the outputs negatively. However, it is important to point out that the number of generated outputs increases compared to the zero-shot approach, reaching the same value as the English baseline.

Finally, the results of the fine-tuned model register an overall increase in performance compared to the few-shot approach. Compared to zero-shot learning, this approach scores lower accuracy and similarity, but registers a higher number of validated potential synonyms.

[7]https://code.google.com/archive/p/word2vec/.
[8]https://zenodo.org/records/8369516 [36].

This is because the number of generated outputs increases greatly, surpassing even the English baseline, which makes accuracy drop since only a portion of the outputs are potential synonyms. While the zero-shot approach is more accurate in output generations, fine-tuning leads to a greater number of generated synonyms and, in turn, of validated potential synonyms. This trade-off might prove advantageous for automating population with a human-in-the-loop approach, since on average a higher number of potential synonyms is generated and the human annotator can efficiently discard inappropriate generations, as the average number of outputs for each input word is moderate (around 5).

Our findings show that the results of the English baseline greatly outperform those of the other approaches across all metrics but the number of generations, which is highest for the fine-tuned model. Considering the progression of the approaches adopted in the experiment, one can note that the scores of accuracy and similarity drop along every stage of the experiment, contrary to the expectations discussed in Section 2.2-2.4, and to the results of [21]. On the other hand, the number of generated outputs steadily increases with each stage of the experiment. The differences in performance across the stages of this experiment, when compared to the results with Latin reported by Santoro et al., are likely due to the language model employed: the model used for this study, Mistral Nemo, is more recent and has a higher number of parameters compared to Mistral 7B, which was used in the study on Latin. The difference in performance between the two models is also reflected in the EB, which scored a much lower accuracy (around 29%, [21]: 4) in Santoro et al.'s work than in the present study (around 90%). Mistral 7B performed poorly in the zero-shot setting, but then registered a marked improvement in the following stages of the experiment. Conversely, Mistral Nemo demonstrated relatively strong performance from the onset, while the few-shot setting scored much lower results, and the fine-tuning led to an increase in potential synonyms, but a decrease in accuracy. Another factor that accounts for the difference in performance between this work and that of Santoro et al.'s is the target language script. It is well documented in the literature that Latin script languages outperform non-Latin script languages across LLM families and in different types of tasks, with a particularly marked disparity in language generation tasks ([37], [38]).

An interesting, yet expected, consideration is that the polysemous dataset outperforms the monosemous dataset across all metrics and approaches but the FS. The results show that the model reaches higher accuracy and similarity scores for the polysemous dataset, generating a greater number of outputs and leading to a higher number of validated potential synonyms. This consideration, which is aligned with the observation and results of [21], applies not only to Ancient Greek, but also to English. A possible explanation for this phenomenon is that polysemous words tend to be more frequent than monosemous words ([39]). As the frequency of a word in pre-training data impacts the LLM's ability to learn its representation ([40]), more frequent words can be linked to higher performance levels, as they are encountered in a wider variety of contexts during model pre-training. Moreover, in a task such as synonym generation, it is likely that language models perform better with polysemous compared to monosemous words, as they encode richer semantic information, resulting in a higher probability of generating suitable outputs. This is because the model is provided with a broader semantic basis from which to draw suitable candidates.

3.2. Qualitative Analysis

Examples of generations across approaches divided for the monosemous and polysemous datasets are shown in Table 3.

Table 3
Examples of generations across approaches. The text not enclosed in parentheses corresponds to the outputs of the model. The lemmas presented in bold type represent validated potential synonyms. The translations provide the meaning of the lemma that justifies the validation as a potential synonym of the target word. Where no translation is provided, the generations are hallucinations of the model, which are presented in roman font.

	Monosemy	Polysemy
Word	*ligús* (shrill)	*krátos* (strength)
ZS	*brakhús* (short), oxûn	***arkhḗ*** (power)
FS	olímos, trílos, fewperos, fewpteros	***hēgemonikón*** (dominant part)
FT	*hēlítēs* (of the sun), ***polús*** (loud)	***dúnamis*** (strength), *pónos* (toil), ***mégethos*** (might), *tíktō*: synonyms: *gígnomai* (generate: synonyms: become), *nosēleúō* (tend a sick person)

One general observation regarding the results is that in all three approaches the model often failed to generate lemmas with the desired PoS. This particular task misalignment also affected the English baseline, even though much less frequently, as in 3:

3 **cumulation**: cumulative.

In this example, despite the mismatch in PoS, the two lemmas share the same root, which is a phenomenon

observed also in some Ancient Greek generations, such as 4.

4 ***homôs*** (similarly): *hómoios* (similar) (FS).

Another type of task misalignment that was frequently observed in Santoro et al. [21] was the generation of multi-word expressions, despite instructions in the prompt explicitly prohibiting it. Notably, such instances are extremely rare in our results, with just a few occurrences (e.g. *met'hautoû* (afterwards) (ZS)).

3.2.1. Non-Ancient Greek Generations

Across all three approaches, the generations include cases of hallucinations, a term that refers to 'generated content that is nonsensical or unfaithful to the provided source content' ([41]). It has been observed in previous literature that hallucinations are amplified by the scarcity of data when dealing with low-resource languages ([42], [43]). Hallucinations are far more frequent in the FS and FT approaches than in ZS. In some cases, the hallucinations share features with the input words, such as the root (see 5.a) or the prefix (5.b). In other cases, no such formal relationship seems to exist (5.c).

5.a ***plêthos*** (multitude): poluplēstía (ZS).

5.b ***diakrínō*** (distinguish): dialúeimi, diēkribállēn (FS).

5.c ***eupetôs*** (easily): tlēmatikós (FT).

Notably, some of the outputs are generated in languages other than Ancient Greek, namely English and Modern Greek, even though the prompt specifically instructs to avoid this behavior (see A.1 and A.2). The inability of LLMs to consistently generate text in a user's desired language is widely known in NLP and is referred to as language confusion ([44]). Examples of language confusion in the model's generations are presented in 6.a and 6.b.

6.a ***arktikós*** (northern): psēlóten/flutter/tall (FT).

6.b ***éris*** (strife): *antagōnismós* (competition) (ZS).

Notably, Mistral models have been found to exhibit high degrees of language confusion ([44]), so the presence of languages other than Ancient Greek in the model's output is not surprising. The problem of English generations also impacted the results of Santoro et al., even though such instances are quite rare in our study. On the contrary, the outputs in Modern Greek are much more numerous, which could depend on an interference effect of the target language's script. This is because the model likely tends to produce outputs in a higher-resource modern language with the same script, as for Latin and English on the one hand, and Ancient Greek and Modern Greek on the other.

3.2.2. Orthographical Errors and Inconsistencies

Taking a closer look at incorrectly generated outputs, several typologies of orthographic errors and inconsistencies were observed. Across approaches, some outputs were written using multiple alphabets: alongside Greek characters, characters from other scripts appeared, such as Latin, Cyrillic, and Arabic (e.g *dapánawm*, *blētérioиы*). Interestingly, these types of errors are less frequent in the zero-shot setting compared to the other approaches.

A second typology of orthographic errors that was observed is closely tied to the internal conventions of Ancient Greek. Across all three training settings, lemmas were generated lacking either the accent (7.a) or the initial breathing mark (7.b). In other cases, the lemmas were generated with an incorrect accent (7.c).

7.a ***krísis*** (dispute): *kindunos* (vs *kíndunos*) (danger) (FT).

7.b ***hellēnikós*** (Greek): *ellēnēios* (vs *hellēnēios*) (Greek) (FS).

7.c ***kritḗs*** (judge): *brabeûs* (vs *brabeús*) (arbiter) (FS).

Notably, such incorrect generations are much less frequent in the zero-shot setting. One may hypothesize that these errors are related to the fact that Modern Greek lacks the initial breathing mark and the iota subscript, and retains a single accent type. A similar type of orthographic inconsistency, affecting only two generations, is the use of the iota adscript instead of the iota subscript. For the target word *kléptēs* (thief), the few-shot and fine-tuning outputs are respectively *lēistḗs* (robber) and *lēïstḗs*. While such instances are linguistically and philologically correct, they were not validated as potential synonyms since they are not compatible with the AGWN graphic standard regarding the iota subscript.

3.2.3. Potential Synonyms

Considering now the generations that were validated as potential synonyms, some interesting observations emerged from the results. One interesting phenomenon that was observed is the generation of rare lemmas or lexical items dating to the Postclassical stages of Ancient Greek (e.g., the Roman or Byzantine period, [45]: 3-6). For example, as a synonym for *kritḗs* (judge) the model generates *lutḗr* (arbitrator), a rare lemma that occurs only 6 times in the Thesaurus Linguae Graecae (TLG)[9]. Only three of such instances are found in Classical texts, while the remaining occurrences come from texts belonging to the Imperial and Byzantine period. Furthermore, the meaning 'arbitrator' associated with *lutḗr* is rare, as it is attested only for one of its occurrences (A.*Th.*940), while

[9]Accessed July, 2025

it usually means 'deliverer'. An example of a generation consisting of a Postclassical lemma is *boreinós* (northern), generated as a synonym for *arktikós* (northern), which is attested 7 times in the TLG, all in Imperial Greek and later, and eventually gives rise to the Modern Greek term *vorinós*. While unexpected, these phenomena do not impact the potential for the automatic population of the AGWN proposed in this work, since the AGWN collects lemmas independently of their frequency or the language stage in which they are attested.

Focusing now on the difference in performance depending on the PoS of the input lemma, Table 4 shows for each approach the number of generations and the number of validated synonyms across PoS, both divided for datasets and overall.

Table 4
Model performance across PoS (Tot: generations for PoS; Syn: potential synonyms for PoS). For each cell, the highest value is presented in bold type to facilitate comparison.

		Overall		**Polysemy**		**Monosemy**	
		Tot	Syn	Tot	Syn	Tot	Syn
ZS	noun	27	9	15	7	12	2
	verb	27	10	15	8	12	2
	adj	**36**	**12**	**19**	**9**	**17**	**3**
	adv	26	3	14	0	12	**3**
FS	noun	40	**6**	14	**2**	26	**4**
	verb	35	2	12	1	23	1
	adj	**54**	0	**23**	0	21	0
	adv	40	1	12	1	**28**	0
FT	noun	**148**	17	107	15	**41**	**2**
	verb	139	**20**	**115**	**18**	24	**2**
	adj	66	5	40	4	26	1
	adv	50	1	26	1	24	0
Total	noun	**215**	**32**	136	24	**79**	**8**
	verb	201	**32**	**142**	**27**	59	5
	adj	156	17	82	13	74	4
	adv	116	5	52	2	64	3

Notably, the PoS for which the model generated the highest number of outputs is nouns (215), followed by verbs (201). However, these overall results are highly influenced by the FT data, which are very abundant and have a great impact on the total. If we consider the ZS and FS approaches alone, the PoS with the most numerous outputs is adjectives (ZS: 36; FS: 54). The PoS with the lowest number of generations is adverbs, a trend that is quite stable across approaches, independently of the dataset considered. Concerning the number of validated synonyms across PoS, the highest number of potential synonyms is generated for nouns (32/215) and verbs (32/201), even though this general trend does not apply to the ZS approach, in which adjectives score the highest number of potential synonyms. Overall, adverbs score the lowest number of potential synonyms (5/116). The reason for this difference in generation trends across PoS may be the distribution of the training data used for fine-tuning, in which nouns and verbs constituted the majority classes, making up, respectively, 54% and 25% of the dataset (see Section 2.1), possibly resulting in a bias of the fine-tuned model. Furthermore, another possible explanation is connected to the difference in performance between the (roughly) polysemous and monosemous datasets already discussed in Section 3.1: independently of the PoS of the input word, the performance of the model is better for polysemous input words across all approaches but FS. Indeed, verbs are generally considered more polysemous than other PoS as their meanings are thought to be more flexible, thus encoding richer semantics ([46], [47]). Nouns also exhibit a high degree of polysemy ([48]). Since, as already discussed, polysemous words tend also to be more frequent, the increase in performance for these PoS may be linked both to a higher frequency in the training data and to their greater polysemy, which provides a broader semantic basis for the generation task at hand.

4. Conclusions

This work has explored the potential of LLMs in the semi-automatic population of the AGWN, evaluating and comparing multiple approaches. The first approach tested was zero-shot, which, despite the lack of examples, generated numerous potential synonyms and achieved considerable accuracy and similarity scores, given the task at hand. Contrary to expectations, the few-shot setting marked a decline in results across all evaluation metrics, except the number of generations. Finally, fine-tuning outperformed the few-shot setting, but scored lower accuracy and similarity values compared to zero-shot prompting. However, this approach scored the highest number of generated outputs and potential synonyms.

The divergence between our results and the outcomes of Santoro et al.'s analysis [21] is likely due to the more recent language model employed, which shows enhanced zero-shot performance, and to the different target language, as the variation in available data and writing system between Greek and Latin can significantly impact the results.

Our analysis shows that, for the task at hand, the zero-shot approach represents a promising starting point for partially automating the population of the AGWN, without needing the resources necessary for fine-tuning a model. Zero-shot generations reach good scores of accuracy and similarity, and in the majority of cases outputs are correctly spelled and lemmatized. On the other hand, while fine-tuning results in lower precision, it leads to a greater number of generations and potential synonyms. This approach, while not as accurate as zero-shot, might prove suitable in a human-in-the-loop scenario, in

which annotators can efficiently discard the inaccurate outputs, accelerating the population process compared to the fewer potential synonyms generated by zero-shot.

The experiments also revealed a marked difference in performance between the two datasets: the model scores higher on the polysemous data across all metrics and approaches, except few-shot. This trend is evident not only in the AG data, but also in the English baseline, and it aligns with the results of Santoro et al. [21]. The explanation for this difference in performance relies on the richer semantic nature of polysemous lemmas, which increases the probability of generating correct outputs. Their increased frequency also positively affects the quality of the representations derived from the model during pre-training.

Closely related to the previous observation is the difference in the number of generated outputs and potential synonyms across PoS. Overall, nouns and verbs score the highest number of generated outputs and potential synonyms, even though there are some variations across approaches (for example, ZS registers the highest number of potential synonyms for adjectives). In contrast, adverbs register the lowest number of generated outputs and potential synonyms, a result which is rather consistent across approaches. These results likely reflect the fact that verbs and nouns constitute the majority classes in the fine-tuning dataset, which probably led to a bias in the model. Furthermore, verbs and nouns are considered highly polysemous PoS, thus the stronger performance on verbs and nouns can be linked to the same factors that lead to better results on the polysemous dataset.

Overall, this study reveals the potential of LLM-based approaches to (partially) automate the annotation of lexical resources. The results, particularly from a qualitative perspective, highlight the specific challenges of working with an ancient and low-resource language such as Ancient Greek. The strategies explored can be used to semi-automatically populate the AGWN by generating candidate synonyms to be validated by a human annotator. This human-in-the-loop approach would significantly reduce the human manual effort, at the same time allowing for a much faster enrichment of the resource.

Acknowledgments

The fine-tuning of the model presented in this work was carried out on the High Performance Computing Data-Center at IUSS, co-funded by Regione Lombardia through the funding programme established by Regional Decree No. 3776 of November 3, 2020. The authors wish to express their sincere gratitude to Cristiano Chesi for granting access to the HPC cluster.

Research for this study was funded through the European Union Funding Program – NextGenerationEU – Missione 4 Istruzione e ricerca - componente 2, investimento 1.1" Fondo per il Programma Nazionale della Ricerca (PNR) e Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN)" progetto PRIN_2022 2022YAPFNJ "Linked WordNets for Ancient Indo-European Languages" CUP F53D2300490 0001 - Dipartimento Studi Umanistici (Università di Pavia) and CUP J53D23008370001 – Dipartimento di Filologia classica, Papirologia e Linguistica storica (Università Cattolica del Sacro Cuore, Milano).

References

[1] C. Fellbaum, Wordnet and wordnets, in: Encyclopedia of Language and Linguistics, Second Edition, Elsevier, 2005.

[2] G. A. Miller, R. Beckwith, C. Fellbaum, D. Gross, K. J. Miller, Introduction to wordnet: An on-line lexical database, International journal of lexicography 3 (1990) 235–244.

[3] C. Fellbaum, WordNet: An electronic lexical database, GMA, MIT Press, 1998.

[4] G. A. Miller, C. Fellbaum, Wordnet then and now, Language Resources and Evaluation 41 (2007) 209–214.

[5] P. Vossen, Introduction to eurowordnet, Computers and the Humanities (1998) 73–89.

[6] S. Minozzi, The latin wordnet project, Latin Linguistics Today. Akten des 15. Internationalem Kol- loquiums zur Lateinischen Linguistik (2009) 707–716.

[7] Y. Bizzoni, F. Boschetti, R. Del Gratta, H. Diakoff, M. Monachini, G. Crane, The making of ancient greek wordnet, Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14) (2014) 1140–1147.

[8] O. Hellwig, The making of ancient greek wordnet, Proceedings of the 12th Inter- national Conference on Computational Semantics (IWCS) 137 (2017) 3934–3941.

[9] E. Biagetti, C. Zanchi, W. M. Short, Toward the creation of WordNets for ancient Indo-European languages, in: P. Vossen, C. Fellbaum (Eds.), Proceedings of the 11th Global Wordnet Conference, Global Wordnet Association, University of South Africa (UNISA), 2021, pp. 258–266. URL: https://aclanthology.org/2021.gwc-1.30/.

[10] F. HKhan, F. J. Minaya Gómez, R. Cruz González, H. Diakoff, J. E. Diaz Vera, J. P. McCrae, C. O'Loughlin, W. M. Short, S. Stolk, Towards the construction of a wordnet for old english, Proceedings of the

Thirteenth Language Resources and Eval- uation Conference, Marseille, France 137 (2022) 3934–3941.
[11] H. G. Liddell, R. Scott, H. S. Jones, R. McKenzie, A Greek–English Lexicon, 9th ed., revised and augmented throughout ed., Clarendon Press, Oxford, 1996.
[12] M. L. Murphy, Lexical meaning, Cambridge University Press, 2010.
[13] E. Sausa, Toward an ancient greek wordnet, ???? Paper presented at the Workshop on WordNet and SketchEngine, Pavia, March 2012.
[14] B. Sagot, D. Fišer, Extending wordnets by learning from multiple resources, in: LTC'11: 5th Language and Technology Conference, 2011.
[15] E. Pianta, L. Bentivogli, C. Girardi, MultiWordNet: developing an aligned multilingual database, in: First International Conference on Global WordNet , 2002.
[16] A. Roventini, A. Alonge, F. Bertagna, N. Calzolari, J. Cancila, C. Girardi, B. Magnini, R. Marinelli, M. Speranza, A. Zampolli, Italwordnet: Building a large semantic database for the automatic treatment of the italian language, Computational Linguistics in Pisa, Special Issue (2003) 745–791.
[17] E. Biagetti, M. Giuliani, S. Zampetta, S. Luraghi, C. Zanchi, Combining neo-structuralist and cognitive approaches to semantics to build wordnets for ancient languages: Challenges and perspectives, in: M. Zock, E. Chersoni, Y.-Y. Hsu, S. de Deyne (Eds.), Proceedings of the Workshop on Cognitive Aspects of the Lexicon @ LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 151–161. URL: https://aclanthology.org/2024.cogalex-1.18/.
[18] P. Singh, G. Rutten, E. Lefever, Pilot study for bert language modelling and morphological analysis for ancient and medieval greek, in: Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, Association for Computational Linguistics, Punta Cana, Dominican Republic (online), 2021, pp. 129–135. URL: https://aclanthology.org/2021.latechclfl-1.15.
[19] J. K. Sandhan, O. Adideva, D. Komal, N. Modani, A. Naik, S. K. Muthiah, M. Kulkarni, Evaluating neural word embeddings for sanskrit, https://arxiv.org/pdf/2104.00270.pdf, 2021. Accessed: [Insert access date here].
[20] A. Mehler, B. Jussen, T. Geelhaar, W. Trautmann, D. Sacha, S. Schwandt, B. Glądalski, D. Lücke, R. Gleim, The frankfurt latin lexicon: From morphological expansion and word embeddings to semiographs, Studi e Saggi Linguistici 58 (2020) 121–155. doi:10.4454/ssl.v58i1.265.
[21] D. Santoro, B. Marchesi, S. Zampetta, M. D. Tredici, E. Biagetti, E. Litta, C. R. Combei, S. Rocchi, T. Facchinetti, R. Ginevra, C. Zanchi, Exploring latin wordnet synset annotation with llms, Global WordNet Conference 2025 54 (2025).
[22] S. C. Woodhouse, English-Greek Dictionary, George Routledge & Sons, Limited, 1910.
[23] V. C. F. Rost, Deutsch-griechisches Wörterbuch, Vandenhöck und Ruprecht, 1829.
[24] F. Montanari, The Brill Dictionary of Ancient Greek, Brill, 2015.
[25] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, Advances in Neural Information Processing Systems (2020).
[26] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, Language models are unsupervised multitask learners | enhanced reader, OpenAI Blog 1 (2019).
[27] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3560815. doi:10.1145/3560815.
[28] L. Reynolds, K. McDonell, Prompt programming for large language models: Beyond the few-shot paradigm, 2021. URL: https://arxiv.org/abs/2102.07350. arXiv:2102.07350.
[29] S. Mishra, D. Khashabi, C. Baral, Y. Choi, H. Hajishirzi, Reframing instructional prompts to gptk's language, 2022. URL: https://arxiv.org/abs/2109.07830. arXiv:2109.07830.
[30] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, Q. V. Le, Finetuned language models are zero-shot learners, in: ICLR 2022 - 10th International Conference on Learning Representations, 2022.
[31] Y. Li, A practical survey on zero-shot prompt design for in-context learning, in: Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings, RANLP, INCOMA Ltd., Shoumen, Bulgaria, 2023, pp. 641–647. URL: http://dx.doi.org/10.26615/978-954-452-092-2_069. doi:10.26615/978-954-452-092-2_069.
[32] A. Radford, Improving language understanding by generative pre-training, Homology, Homotopy and Applications 9 (2018).
[33] E. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li,

S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: ICLR 2022 - 10th International Conference on Learning Representations, 2022.
[34] J. Myerston, J. López, grecy: Ancient greek spacy models for natural language processing in python, 2023.
[35] J. L. Fleiss, Measuring nominal scale agreement among many raters, Psychological Bulletin 76 (1971). doi:10.1037/h0031619.
[36] S. Stopponi, N. Pedrazzini, S. Peels-Matthey, B. McGillivray, M. Nissim, Natural language processing for ancient greek, Diachronica 41 (2024) 414–435. URL: https://www.jbe-platform.com/content/journals/10.1075/dia.23013.sto. doi:https://doi.org/10.1075/dia.23013.sto.
[37] H. Nguyen, K. Mahajan, V. Yadav, J. Salazar, P. S. Yu, M. Hashemi, R. Maheshwary, Prompting with phonemes: Enhancing llms' multilinguality for non-latin script languages, in: L. Chiruzzo, A. Ritter, L. Wang (Eds.), Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), Association for Computational Linguistics, 2025, pp. 11975–11994. URL: https://aclanthology.org/2025.naacl-long.599/. doi:10.18653/v1/2025.naacl-long.599.
[38] O. Shliazhko, A. Fenogenova, M. Tikhonova, A. Kozlova, V. Mikhailov, T. Shavrina, mgpt: Few-shot learners go multilingual, Transactions of the Association for Computational Linguistics 12 (2024). doi:10.1162/tacl_a_00633.
[39] G. K. Zipf, The meaning-frequency relationship of words, Journal of General Psychology 33 (1945). doi:10.1080/00221309.1945.10544509.
[40] T. Fu, R. Ferrando, J. Conde, C. Arriaga-Prieto, P. Reviriego, Why do large language models (llms) struggle to count letters?, 2024. doi:10.48550/arXiv.2412.18626.
[41] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55 (2023). URL: https://doi.org/10.1145/3571730. doi:10.1145/3571730.
[42] N. M. Guerreiro, D. M. Alves, J. Waldendorf, B. Haddow, A. Birch, P. Colombo, A. F. Martins, Hallucinations in large multilingual translation models, Transactions of the Association for Computational Linguistics 11 (2023). doi:10.1162/tacl_a_00615.
[43] M. Abdelrahman, Hallucination in low-resource languages: Amplified risks and mitigation strategies for multilingual llms, Journal of Applied Big Data Analytics, Decision-Making, and Predictive Modelling Systems 8 (2024) 17–24. URL: https://polarpublications.com/index.php/JABADP/article/view/2024-12-10.
[44] K. Marchisio, W.-Y. Ko, A. Bérard, T. Dehaze, S. Ruder, Understanding and mitigating language confusion in llms, 2024. doi:10.48550/arXiv.2406.20052.
[45] G. D. Bartolo, D. Kölligan, Postclassical Greek: Problems and Perspectives, De Gruyter, 2024.
[46] C. Fellbaum, English Verbs as a Semantic Net, International Journal of Lexicography 3 (1990) 278–301. URL: http://dx.doi.org/10.1093/ijl/3.4.278. doi:10.1093/ijl/3.4.278.
[47] D. Gentner, I. M. France, The verb mutability effect: Studies of the combinatorial semantics of nouns and verbs, 2013. doi:10.1016/B978-0-08-051013-2.50018-5.
[48] A. A. Freihat, F. Giunchiglia, B. Dutta, A taxonomic classification of wordnet polysemy types, in: Proceedings of the 8th Global WordNet Conference, GWC 2016, 2016.

Online Resources

- Thesaurus Linguae Graecae® Digital Library. Ed. Maria C. Pantelia. University of California, Irvine (accessed May 31 2025).

A. Prompts Used in the Experiment

This appendix contains the full prompts used in the experiment for both Ancient Greek and English.

A.1. Ancient Greek Prompt

```
 zs_prompt = f"""You are a powerful
AI assistant trained in semantics and
Classics.
You are an Ancient Greek native
speaker. The only language you speak
is Ancient Greek.
Your task is to provide a bullet list
of Ancient Greek synonyms for a user-
chosen word.
Your response must contain the
generated synonyms as comma-separated
values.
Observe the following instructions
very closely: [INST]
- Generate only Ancient Greek
synonyms.
```

```
- Provide single-word expressions
ONLY.
- Do NOT generate long phrases.
- Make sure to provide numerous
synonyms for each lemma.
-- ABSOLUTELY AVOID including any
additional explanations or comments
in your output.
- VERY IMPORTANT: DO NOT translate the
words.
- VERY IMPORTANT: Use ANCIENT GREEK
exclusively.
- VERY IMPORTANT: Generate ANCIENT
GREEK lemmas in the original script
with accurate diacritics (accents,
breathing marks, and vowel quantity
for long vowels indicated by macrons
or other notations).
- VERY IMPORTANT: Make sure the
outputs are spelled correctly.
- IMPORTANT: Do NOT generate any word
in Modern Greek.
- IMPORTANT: Generate words with the
same part of speech as the input
word,
for example if the input word is a
verb you must generate only verbs as
synonyms.
-- For NOUNS generate only the
NOMINATIVE CASE, as shown in the
examples below.
-- For VERBS generate only the FIRST-
PERSON SINGULAR of the INDICATIVE.
-- List each Ancient Greek word
separately with proper formatting.
"""
```

A.2. English Prompt

```
 en_prompt=f"""You are a powerful AI
assistant trained in semantics. You
are an English native speaker. Your
task is to provide a bullet list of
English synonyms for a user-chosen
word.
Your response must contain the
generated synonyms as comma-separated
values.
Observe the following instructions
very closely: [INST]
- Generate only English synonyms.
- Provide single-word expressions
ONLY.
- Do NOT generate long phrases.
- Make sure to provide numerous
synonyms for each lemma.
-- ABSOLUTELY AVOID including any
additional explanations or comments
in your output.
- VERY IMPORTANT: Make sure the
outputs are spelled correctly.
- IMPORTANT: Generate words with the
same part of speech as the input word,
for example if the input word is a
verb you must generate only verbs as
synonyms.
-- List each English word separately
with proper formatting.
"""
```

A.3. Examples for the Few-Shot Prompt

word: 'nouthetḗseis'
synonyms: ['paramuthía', 'protropḗ', 'parakéleusis', 'parórmēsis', 'paroksusmós', 'peithṓ', 'pístis', 'kéntron', 'múōps', 'paraínesis']

word: 'atimázō'
synonyms: ['kataiskhúnō', 'aischúnō', 'atimóō', 'atimáō']

word: 'theosebḗs'
synonyms: ['deisidaímōn', 'eusebēḗs', 'eúphēmos', 'pistós']

word: 'autoû'
synonyms: ['entaûtha', 'enthádе', 'autóthi', 'éntha', 'ekeî']

word: 'trophḗ'
synonyms: ['deîpnon', 'edōdḗ', 'sîtos', 'édesma']

word: 'elassóō'
synonyms: ['koloúō', 'meióō', 'tapeinóō', 'aphairéō', 'diaphtheírō']

word: 'iskhurós'
synonyms: ['drastḗrios', 'karterós', 'energḗs', 'rhōmaléos', 'krataíos', 'óbrimos', 'sthenarós', 'kraterós']

word: 'oknērôs'
synonyms: ['phoberôs', 'deilôs']

Declaration on Generative AI

During the preparation of this work, the author(s) did not use any generative AI tools or services.

Author Index