

CLiC-it 2025

**The Eleventh Italian Conference on Computational  
Linguistics**

**Proceedings of the Conference**

September 24-26, 2025

Copyright ©2025 for the individual papers by the papers' authors.  
Copyright ©2025 for the volume as a collection by its editors.  
This volume and its papers are published under the Creative Commons License Attribution 4.0 International (CC BY 4.0).

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*

Cristina Bosco, Elisabetta Jezek, Marco Polignano and Manuela Sanguinetti ..... 1

## *Contemporary Voices in Ancient Tongue: Integrating Papal Encyclicals into the LiLa KB*

Aurora Alagni, Federica Iurescia and Eleonora Litta ..... 4

## *Ellipsis in Enhanced Dependencies: A Case Study on Latin*

Lisa Sophie Albertelli, Lorenzo Augello, Giulia Calvi, Annachiara Clementelli, Federica Iurescia and Claudia Corbetta ..... 12

## *Low- vs High-level Lemmatization for Historical Languages. a Case Study on Italian*

Chiara Alzetta and Simonetta Montemagni ..... 21

## *PeRAG: Multi-Modal Perspective-Oriented Verbalization with RAG for Inclusive Decision Making*

Muhammad Saad Amin, Horacio Jesús Jarquín-Vásquez, Franco Sansonetti, Simona Lo Giudice, Valerio Basile and Viviana Patti ..... 31

## *Beyond Raw Text: Knowledge-Augmented Italian Relation Extraction with Large Language Models*

Gianmaria Balducci, Elisabetta Fersini and Messina Enza ..... 45

## *When Figures Speak with Irony: Investigating the Role of Rhetorical Figures in Irony Generation with LLMs*

Pier Felice Balestrucci, Michael Oliverio, Soda Maren Lo, Luca Anselma, Valerio Basile, Alessandro Mazzei and Viviana Patti ..... 55

## *BLiMP-IT: Harnessing Automatic Minimal Pair Generation for Italian Language Model Evaluation*

Matilde Barbini, Maria Letizia Piccini Bianchessi, Veronica Bressan, Achille Fusco, Sofia Neri, Sarah Rossi, Tommaso Sgrizzi and Cristiano Chesi ..... 64

## *Do LLMs Authentically Represent Affective Experiences of People with Disabilities on Social Media?*

Marco Bombieri, Simone Paolo Ponzetto and Marco Rospocher ..... 72

## *LLMs Struggle on Explicit Causality in Italian*

Alessandro Bondielli, Martina Miliani, Luca Paglione, Serena Auriemma, Lucia C. Passaro and Alessandro Lenci ..... 83

## *Sparse Autoencoders Find Partially Interpretable Features in Italian Small Language Models*

Alessandro Bondielli, Lucia C. Passaro and Alessandro Lenci ..... 95

## *A Novel Real-World Dataset of Italian Clinical Notes for NLP-based Decision Support in Low Back Pain Treatment*

Agnese Bonfigli, Ruben Piperno, Luca Bacco, Felice Dell’Orletta, Dominique Brunato, Filippo Crispino, Giuseppe Francesco Papalia, Fabrizio Russo, Gianluca Vadalà, Rocco Papalia, Mario Merone and Leandro Pecchia ..... 105

## *MakeItSample: A Python Library for Generating Typological Language Samples Based on the Diversity Value Metric*

Luca Brigada Villa ..... 116

## *The OuLiBench Benchmark: Formal Constraints as a Lens into LLM Linguistic Competence*

Silvio Calderaro, Alessio Miaschi and Felice Dell’Orletta ..... 124

## *BES4RAG: A Framework for Embedding Model Selection in Retrieval-Augmented Generation*

Lorenzo Canale, Stefano Scotta, Alberto Messina and Laura Farinetti ..... 134

<i>BAMBI Goes to School: Evaluating Italian BabyLMs with Invalsi-ITA</i> Luca Capone, Alice Suozzi, Gianluca Leboni and Alessandro Lenci .....	143
<i>Arbuli Sunnu: A Sicilian-Italian Parallel Treebank</i> Caterina Maria Cappello, Sabrina D’Alì, Mario Guglielmetti, Elisa Di Nuovo and Cristina Bosco	155
<i>A BERT-Based Approach for Part-of-Speech Tagging in the Low-Resource Context of Sardinian</i> Salvatore Mario Carta, Filippo Concas, Gianni Fenu, Alessandro Giuliani, Marco Manolo Manca, Mirko Marras, Piergiorgio Mura and Simone Pisano .....	169
<i>A Hypothesis-Driven Framework for Detecting Lexical Semantic Change</i> Pierluigi Cassotti and Nina Tahmasebi .....	177
<i>Balancing Translation Quality and Environmental Impact: Comparing Large and Small Language Models</i> Antonio Castaldo, Petra Giommarelli and Johanna Monti .....	186
<i>On the Impact of Hate Speech Synthetic Data on Model Fairness</i> Camilla Casula and Sara Tonelli .....	194
<i>Classifying Gas Pipe Damage Descriptions in Low-Diversity Corpora</i> Luca Catalano, Federico D’Asaro, Michele Pantaleo, Minal Jamshed, Prima Acharjee, Nicola Giulietti, Eugenio Fossat and Giuseppe Rizzo .....	206
<i>Knowledge-Grounded Detection of Factual Hallucinations in Large Language Models</i> Cristian Ceccarelli, Alessandro Raganato and Marco Viviani .....	216
<i>Benchmarking Historical Phase Recognition from Text and Events</i> Fabio Celli and Marco Rovera .....	225
<i>Ontology-Guided Domain Entity Recognition in Environmental Texts: Evaluating Syntax-Driven and LLM Approaches Using BabelNet and GEMET</i> Elisa Chierchiello, Patricia Chiril and Adriana Pagano .....	233
<i>Crossword Space: Latent Manifold Learning for Italian Crosswords and beyond</i> Cristiano Ciaccio, Gabriele Sarti, Alessio Miaschi and Felice Dell’Orletta .....	245
<i>Veras Audire Et Reddere Voces: A Corpus of Prosodically-Correct Latin Poetic Audio from Large-Language-Model TTS</i> Michele Ciletti .....	256
<i>A Tough Hoe to Row: Instruction Fine-Tuning LLaMA 3.2 for Multilingual Idiom Processing</i> Debora Ciminari and Alberto Barrón-Cedeño .....	263
<i>Semantic Priming in GPT: Investigating LLMs through a Cognitive Psychology Lens</i> Filippo Colombi and Carlo Strapparava .....	273
<i>Verso La Valutazione Automatizzata Dell’italiano L2: ETET Tra LLM E Tecnologie Vocali</i> Anna Vignoli, Claudia Roberta Combei and Francesco Zappulla .....	282
<i>What Is Better for Syntactic Parsing? A Comparison between Supervised and Unsupervised Models on Dante and Cavalcanti</i> Claudia Corbetta, Anna Erminia Colombi, Giovanni Moretti and Marco Passarotti .....	292
<i>Segmenting Italian Sentences for Easy Reading</i> Marta Cozzini and Horacio Saggion .....	304

<i>Unveiling Stereotypes: Combining Knowledge Graphs and LLMs for Implied Stereotype Generation</i> Marco Cuccarini, Lia Draetta, Beatrice Fiumanò, Stefano Bistarelli, Rossana Damiano and Valentina Presutti . . . . .	314
<i>Detecting Semantic Reuse in Ancient Greek Literature: A Computational Approach</i> Caterina D’Angelo, Andrea Taddei and Alessandro Lenci . . . . .	327
<i>WorthIt: Check-worthiness Estimation of Italian Social Media Posts</i> Agnese Daffara, Alan Ramponi and Sara Tonelli . . . . .	337
<i>Diffusion-Aided RAG: Elevating Dense-Retrieval Chatbots via Graph-Based Diffusion Reranking</i> Sai Teja Dampanaboina, Sai Nishchal Gamini, Karishma Kunwar, Marco Polignano, Marco Levantesi, Giovanni Semeraro and Ernesto William De Luca . . . . .	352
<i>When Less Is More? Diagnosing ASR Predictions in Sardinian via Layer-Wise Decoding</i> Domenico De Cristofaro, Alessandro Vietti, Marianne Pouplier and Aleese Block . . . . .	363
<i>Meta-evaluation of Automatic Machine Translation Metrics between Italian and a Minor Language Variety of German</i> Paolo Di Natale, Elena Chiochetti and Egon Waldemar Stemle . . . . .	371
<i>Linking Emotions: Affective and Lexical Resources for Italian in Linked Open Data</i> Eliana Di Palma, Valerio Basile, Agnese Vardanega, Giuliano Gabrieli and Marco Vassallo . . . . .	384
<i>Dissonant Ballerinas and Crafty Carrots: A Comparative Multi-modal Analysis of Italian Brain Rot</i> Anca Dinu, Andra-Maria Florescu, Marius Micluța-Câmpeanu, Ștefana Arina Tăbușcă, Claudiu Creangă and Andreiana Mihail . . . . .	397
<i>The Role of Eye-Tracking Data in Encoder-Based Models: An In-depth Linguistic Analysis</i> Lucia Domenichelli, Luca Dini, Dominique Brunato and Felice Dell’Orletta . . . . .	408
<i>Seeing Cause and Time: A Visually Grounded Evaluation of Multimodal Models</i> Salvatore Ergoli, Alessandro Bondielli and Alessandro Lenci . . . . .	423
<i>Mapping Meaning in Latin with Large Language Models: A Multi-Task Evaluation of Preverbed Motion Verbs and Spatial Relation Detection in LLMs</i> Andrea Farina, Andrea Ballatore and Barbara McGillivray . . . . .	434
<i>MAMITA: Benchmarking Misogyny in Italian Memes</i> Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi and Aurora Saibene . . . . .	445
<i>LEARN: On the Feasibility of Learner Error AutoRegressive Neural Annotation</i> Paolo Gajo, Daniele Polizzi, Adriano Ferraresi and Alberto Barrón-Cedeño . . . . .	456
<i>The Meaning of Beatus: Disambiguating Latin with Contemporary AI Models</i> Eleonora Ghizzota, Pierpaolo Basile, Lucia Siciliani and Giovanni Semeraro . . . . .	469
<i>LLMike: Exploring Large Language Models’ Abilities in Wheel of Fortune Riddles</i> Ejdis Gjinić, Nicola Arici, Andrea Loreggia, Luca Putelli, Ivan Serina and Alfonso Emilio Gevini . . . . .	480
<i>Bidirectional Emotional Influence in Human-LLM Interaction: Empirical Analysis and Methodological Framework</i> Manuel Gozzi and Francesca Fallucchi . . . . .	490
<i>A Multilingual Investigation of Anthropocentrism in GPT-4O</i> Francesca Grasso and Stefano Locci . . . . .	500

<i>Surprisal and Crossword Clues Difficulty: Evaluating Linguistic Processing between LLMs and Humans</i>	
Tommaso Iaquinta, Asya Zanollo, Achille Fusco, Kamyar Zeinalipour and Cristiano Chesi . .	512
<i>AI-Driven Resume Analysis and Enhancement Using Semantic Modeling and Large Language Feedback Loops</i>	
Achal Jagadeesh, Chinmayi Ravi Shankar, Sahithya Narayanaswamy Patel, Marco Levantesi, Giovanni Semeraro and Ernesto William De Luca . . . . .	526
<i>DIETA: A Decoder-only Transformer-based Model for Italian-English Machine Translation</i>	
Pranav Kasela, Marco Braga, Alessandro Ghiotto, Andrea Pilzer, Marco Viviani and Alessandro Raganato . . . . .	539
<i>Positional Bias in Binary Question Answering: How Uncertainty Shapes Model Preferences</i>	
Tiziano Labruna, Simone Gallo and Giovanni Da San Martino . . . . .	550
<i>Is It Still a Village? Tracing Grammaticalization with Word Embeddings</i>	
Joseph E. Larson and Patrícia Amaral . . . . .	561
<i>MedBench-IT: A Comprehensive Benchmark for Evaluating Large Language Models on Italian Medical Entrance Examinations</i>	
Ruggero Marino Lazzaroni, Alessandro Angioi, Michelangelo Puliga, Davide Sanna and Roberto Marras . . . . .	570
<i>MuLTa-Telegram: A Fine-Grained Italian and Polish Dataset for Hate Speech and Target Detection</i>	
Elisa Leonardelli, Camilla Casula, Sebastiano Vecellio Salto, Joanna Ewa Bak, Elisa Muratore, Anna Kofos, Thomas Louf and Sara Tonelli . . . . .	582
<i>Linking CompL-it to the LiITA Knowledge Base</i>	
Eleonora Litta, Marco Passarotti, Giovanni Moretti, Paolo Brasolin, Francesco Mambrini, Valerio Basile, Andrea Di Fabio, Eliana Di Palma, Emiliano Giovannetti, Simone Marchi, Andrea Bellandi and Flavia Sciolette . . . . .	595
<i>Subjectivity in Stereotypes against Migrants in Italian: An Experimental Annotation Procedure</i>	
Soda Marem Lo, Marco Antonio Stranisci, Alessandra Teresa Cignarella, Simona Frenda, Valerio Basile, Elisabetta Jezek and Viviana Patti . . . . .	603
<i>Doing Things with Words: Rethinking Theory of Mind Simulation in Large Language Models</i>	
Agnese Lombardi and Alessandro Lenci . . . . .	613
<i>BeaverTails-IT: Towards a Safety Benchmark for Evaluating Italian Large Language Models</i>	
Giuseppe Magazzù, Alberto Sormani, Giulia Rizzi, Francesca Pulerà, Daniel Scalena, Stefano Cariddi, Edoardo Michielon, Marco Pasqualini, Claudio Stamile and Elisabetta Fersini . . . . .	625
<i>A Leaderboard for Benchmarking LLMs on Italian</i>	
Bernardo Magnini, Marco Madeddu, Michele Resta, Roberto Zanolli, Martin Cimmino, Paolo Albano and Viviana Patti . . . . .	636
<i>Towards the Semi-Automated Population of the Ancient Greek WordNet</i>	
Beatrice Marchesi, Annachiara Clementelli, Andrea Maurizio Mammarella, Silvia Zampetta, Erica Biagetti, Luca Brigada Villa, Virginia Mastellari, Riccardo Ginevra, Claudia Roberta Combei and Chiara Zanchi . . . . .	647
<i>Evaluating Large Language Models on Wikipedia Graph Navigation: Insights from the WikiGame</i>	
Daniele Margiotta, Danilo Croce and Roberto Basili . . . . .	659

<i>Linguistic Markers of Population Replacement Conspiracy Theories in YouTube Immigration Discourse</i> Erik Bran Marino, Davide Bassi and Renata Vieira .....	670
<i>Exploring the Adaptability of Large Speech Models to Non-Verbal Vocalization Task</i> Juan José Márquez Villacís, Federico D’Asaro, Giuseppe Rizzo and Andrea Bottino .....	680
<i>Strategic Conversations: LLMs Argumentation and User Perception in Movie Recommendation Dialogues</i> Valeria Mauro, Martina Di Bratto, Valentina Russo, Azzurra Mancini and Marco Grazioso ..	689
<i>Uni-Mate: A Retrieval-Augmented Generation System to Provide High School Students with Accurate Academic Guidance</i> Samuele Mazzei, Lorenzo Zambotto, Gabriele Tealdo, Alberto Macagno and Alessio Palmero Aprosio .....	699
<i>Language Models and the Magic of Metaphor: A Comparative Evaluation with Human Judgments</i> Simone Mazzoli, Alice Suozzi and Gianluca Lebani .....	710
<i>Easy to Complete, Hard to Choose: Investigating LLM Performance on the ProverbIT Benchmark</i> Enrico Mensa, Lorenzo Zane, Calogero Jerik Scozzaro, Matteo Delsanto, Tommaso Milani and Daniele P. Radicioni .....	722
<i>Mamma Mia! Where’s My Name? De-Identifying Italian Clinical Notes with Large Language Models</i> Michele Miranda, Sébastien Bratières, Stefano Patarnello and Livia Lilli .....	735
<i>Sustainable Italian LLM Evaluation: Community Perspectives and Methodological Guidelines</i> Luca Moroni, Gianmarco Pappacoda, Edoardo Barba, Simone Conia, Andrea Galassi, Bernardo Magnini, Roberto Navigli, Paolo Torroni and Roberto Zanoli .....	747
<i>What We Learned from Continually Training Minerva: A Case Study on Italian</i> Luca Moroni, Tommaso Bonomo, Luca Gioffré, Lu Xu, Domenico Fedele, Leonardo Colosi, Andrei Stefan Bejgu, Alessandro Scirè and Roberto Navigli .....	760
<i>Automatic GRI-SDG Annotation and LLM-Based Filtering for Sustainability Reports</i> Seyed Alireza Mousavian Anaraki, Danilo Croce and Roberto Basili .....	775
<i>Benchmarking Large Language Models for Target-Based Financial Sentiment Analysis</i> Iftikhar Muhammad, Marco Rospocher, Timotej Knez and Slavko Žitnik .....	785
<i>Is Multimodality Still Required for Multimodal Machine Translation? A Case Study on English and Italian</i> Elio Musacchio, Lucia Siciliani, Pierpaolo Basile and Giovanni Semeraro .....	796
<i>Extending Italian Large Language Models for Vision-language Tasks</i> Elio Musacchio, Lucia Siciliani, Pierpaolo Basile, Asia Beatrice Uboldi, Giovanni Germani and Giovanni Semeraro .....	805
<i>Probing Feminist Representations: A Study of Bias in LLMs and Word Embeddings</i> Arianna Muti, Elisa Bassignana, Emanuele Moscato and Debora Nozza .....	815
<i>Multilingual vs. Monolingual Transformer Models in Encoding Linguistic Structure and Lexical Abstraction</i> Vivi Nastase, Giuseppe Samo, Chunyang Jiang and Paola Merlo .....	826
<i>Direct and Indirect Interpretations of Speech Acts: Evidence from Human Judgments and Large Language Models</i> Massimiliano Orsini and Dominique Brunato .....	837

<i>Narrative Conflicts: A Tri-Modal Computational Analysis of Antagonism in Shakespeare’s Julius Caesar</i>	
Aylin Özkan, İrem Ertuş, Mehmet Can Yavuz and Lucia Cascone	849
<i>FAMA: The First Large-Scale Open-Science Speech Foundation Model for English and Italian</i>	
Sara Papi, Marco Gaido, Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, Roberto Gretter, Marco Matassoni, Mohamed Nabih Ali Mohamed Nawar and Matteo Negri	860
<i>Generating and Evaluating Multi-Level Text Simplification: A Case Study on Italian</i>	
Michele Papucci, Giulia Venturi and Felice Dell’Orletta	870
<i>Analyzing Femicide Reactions in YouTube Comments: A Comparative Study of Giulia Cecchettin and Carol Maltesi</i>	
Sveva Silvia Pasini, Marco Madeddu, Chiara Ferrando, Chiara Zanchi and Viviana Patti	886
<i>Gender-Neutral Rewriting in Italian: Models, Approaches, and Trade-offs</i>	
Andrea Piergentili, Beatrice Savoldi, Matteo Negri and Luisa Bentivogli	899
<i>Doctor, Is That You? Evaluating Large Language Models on Italy’s Medical School Entrance Exams</i>	
Ruben Piperno, Agnese Bonfigli, Felice Dell’Orletta, Leandro Pecchia, Mario Merone and Luca Bacco	911
<i>A Modular LLM-based Dialog System for Accessible Exploration of Finite State Automata</i>	
Stefano Vittorio Porta, Pier Felice Balestrucci, Michael Oliverio, Luca Anselma and Alessandro Mazzei	924
<i>Leveraging LLMs to Build a Semi-synthetic Dataset for Legal Information Retrieval: A Case Study on the Italian Civil Code and GPT4-O</i>	
Mattia Proietti, Lucia C. Passaro and Alessandro Lenci	933
<i>No Longer Left Behind: Self-training Reasoning Models in Italian</i>	
Federico Ranaldi and Leonardo Ranaldi	942
<i>Evaluating Models, Prompting Strategies, and Task Formats: A Case Study on the MACID Challenge</i>	
Matteo Rinaldi, Rossella Varvara, Lorenzo Gregori and Andrea Amelio Ravelli	955
<i>Gender Violence in Numbers: Prompting Italian LLMs to Characterize Crimes against Women</i>	
Giulia Rizzi, Daniel Scalena and Elisabetta Fersini	965
<i>Uncovering Unsafety Traits in Italian Language Models</i>	
Giulia Rizzi, Giuseppe Magazzù, Alberto Sormani, Francesca Pulerà, Daniel Scalena and Elisabetta Fersini	974
<i>Can LLMs Help Recollect and Elaborate on Our Personal Experiences?</i>	
Gabriel Roccabruna, Olha Khomyn, Michele Yin and Giuseppe Riccardi	983
<i>IMB: An Italian Medical Benchmark for Question Answering</i>	
Antonio Romano, Giuseppe Riccio, Mariano Barone, Marco Postiglione and Vincenzo Moscato	995
<i>Acquisition in Babies and Machines: Comparing the Learning Trajectories of LMs in Terms of Syntactic Structures (ATTracTSS Test Set)</i>	
Sarah Rossi, Guido Formichi, Sofia Neri, Tommaso Sgrizzi, Asya Zanollo, Veronica Bressan and Cristiano Chesi	1007
<i>Context-Aware Search Space Adaptation of Hyperparameters and Architectures for AutoML in Text Classification</i>	
Parisa Safikhani and David Broneske	1018

<i>Evaluating Linguistic Speaker Profiles on Response Selection in Multi-Party Dialogue</i>	
Maryam Sajedinia, Seyed Mahed Mousavi and Valerio Basile .....	1028
<i>MLLMs Construction Company: Investigating Multimodal LLMs' Communicative Skills in a Collaborative Building Task</i>	
Marika Sarzotti, Giovanni Duca, Chris Madge, Raffaella Bernardi and Massimo Poesio ....	1037
<i>Toward Optimised Datasets to Fine-tune ASR Systems Leveraging Less but More Informative Speech</i>	
Loredana Schettino, Vincenzo Norman Vitale and Alessandro Vietti .....	1048
<i>Using End-to-End Automatic Speech Recognizers' Internals to Model Disfluencies in Italian Patients with Early-stage Parkinson's Disease</i>	
Loredana Schettino, Vincenzo Norman Vitale and Marta Maffia .....	1055
<i>Structural Sensitivity Does Not Entail Grammaticality: Assessing LLMs against the Universal Functional Hierarchy</i>	
Tommaso Sgrizzi, Asya Zanollo and Cristiano Chesi .....	1063
<i>Evaluation of Italian and English Small Language Models for Domain-based QA in Low-Resource Scenario</i>	
Irene Siragusa and Roberto Pirrone .....	1074
<i>Annotating Manzoni: Challenges in the Annotation of Lemmas, POS and Features in "I Promessi Sposi"</i>	
Rachele Sprugnoli and Arianna Redaelli .....	1084
<i>Ciallabacialla! Modeling and Linking a Regional Lexical Resource to Include Sicilian in the Semantic Web</i>	
Rachele Sprugnoli, Giovanni Moretti, Domenico Giuseppe Muscianisi and Eleonora Litta ..	1093
<i>Curated Data Does Not Mean Representative Data When Training Large Language Models: An Experiment Using Representative Data for Italian</i>	
Fabio Tamburini .....	1102
<i>BABILong-ITA: A New Benchmark for Testing Large Language Models Effective Context Length and a Context Extension Method</i>	
Fabio Tamburini .....	1112
<i>MAIA: A Benchmark for Multimodal AI Assessment</i>	
Davide Testa, Giovanni Bonetta, Raffaella Bernardi, Alessandro Bondielli, Alessandro Lenci, Alessio Miaschi, Lucia C. Passaro and Bernardo Magnini .....	1121
<i>Building It-tok: An Italian TikTok Corpus</i>	
Luisa Troncone .....	1135
<i>Protecting the Privacy in Velvet with Model Editing</i>	
Giancarlo A. Xompero, Elena Sofia Ruzzetti, Cristina Giannone, Andrea Favalli, Raniero Romagnoli and Fabio Massimo Zanzotto .....	1148
<i>I Understand, but...": Towards a Comprehensive Account of the Explainee's Voice in Explanatory Dialogues</i>	
Andrea Zaninello, Petar Bodlovic, Marcin Lewinski and Bernardo Magnini .....	1158
<i>PharmaER.IT: An Italian Dataset for Entity Recognition in the Pharmaceutical Domain</i>	
Andrea Zugarini and Leonardo Rigutini .....	1171

# Towards the Semi-Automated Population of the Ancient Greek WordNet

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

<sup>1</sup>Università degli Studi di Pavia

<sup>2</sup>Università Cattolica del Sacro Cuore

<sup>3</sup>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*, *phōtér*, *apaúgasma*, *periphéggeia*, *augasmós*, *bolé*, *kiállē*<sup>1</sup> 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<sup>2</sup>. 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)

© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

<sup>1</sup>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].

<sup>2</sup>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].)

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<sup>3</sup> 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<sup>4</sup>, 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

<sup>3</sup>The datasets, code, and data used for this experiment are provided in a repository at <https://github.com/unipv-larl/llms-ag>.

<sup>4</sup><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, aitiama, tó, énklēma, tó, V. epiklēma, 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:

- **phrikódēs** (awe-inspiring): *ouránios* (heavenly), *theios* (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%)<sup>5</sup>.

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

<sup>5</sup>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
<b>Training loss</b>	1,2943	1,4099	<b>1,1478</b>	1,2232	1,1855
<b>Validation loss</b>	1,4814	1,4366	1,4137	<b>1,4087</b>	1,4100
<b>Training mean token accuracy</b>	0,6587	0,6262	0,6597	0,6720	<b>0,7206</b>

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<sup>6</sup> validated the results, determining for each generated item if it constituted a potential synonym of the input word. In

<sup>6</sup>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.  
*kratēō*: 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	<b>30%</b>	<b>.261</b>	116	34
	FS	5%	.099	169	9
	FT	11%	.077	<b>403</b>	<b>43</b>
Polysemy	EB	98%	.407	85	83
	ZS	<b>40%</b>	<b>.296</b>	63	24
	FS	7%	.066	61	4
	FT	13%	.113	<b>288</b>	<b>38</b>
Monosemy	EB	83%	.347	82	68
	ZS	<b>19%</b>	<b>.226</b>	53	<b>10</b>
	FS	5%	.132	108	5
	FT	4%	.041	<b>115</b>	5

As for the similarity metric, cosine similarity was computed using pre-trained Word2vec embeddings based on a skip-Gram model for both English<sup>7</sup> and Ancient Greek<sup>8</sup>. 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.

<sup>7</sup><https://code.google.com/archive/p/word2vec/>.

<sup>8</sup><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	<i>ligús</i> (shrill)	<i>krátos</i> (strength)
ZS	<i>brakhús</i> (short), <i>oxûn</i>	<b><i>arkhế</i></b> (power)
FS	<i>olímos</i> , <i>trílos</i> , fewperos, fewpteros	<b><i>hēgemonikón</i></b> (dominant part)
FT	<i>hēlítēs</i> (of the sun), <b><i>polús</i></b> (loud)	<b><i>dúnamis</i></b> (strength), <i>pónos</i> (toil), <b><i>mégethos</i></b> (might), <i>tiktō</i> : synonyms: <i>gígnomai</i> (generate: synonyms: become), <i>nosēleúō</i> (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 *homōs* (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'hautoú* (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 *plēthos* (multitude): *poluplēstía* (ZS).

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

5.c *eupetōs* (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): *psēlótēn/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. *dapána<sup>wm</sup>, blētéri<sup>on</sup>bl*). 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 *kritēs* (judge): *brabeús* (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 *lēistēs* (robber) and *lēístēs*. 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 *kritēs* (judge) the model generates *lutér* (arbitrator), a rare lemma that occurs only 6 times in the Thesaurus Linguae Graecae (TLG)<sup>9</sup>. 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 *lutér* is rare, as it is attested only for one of its occurrences (A.Th.940), while

<sup>9</sup>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 *arctikós* (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	<b>36</b>	<b>12</b>	<b>19</b>	<b>9</b>	<b>17</b>	<b>3</b>
	adv	26	3	14	0	12	<b>3</b>
FS	noun	40	<b>6</b>	14	<b>2</b>	26	<b>4</b>
	verb	35	2	12	1	23	1
	adj	<b>54</b>	0	<b>23</b>	0	21	0
	adv	40	1	12	1	<b>28</b>	0
FT	noun	<b>148</b>	17	107	15	<b>41</b>	<b>2</b>
	verb	139	<b>20</b>	<b>115</b>	<b>18</b>	24	<b>2</b>
	adj	66	5	40	4	26	1
	adv	50	1	26	1	24	0
Total	noun	<b>215</b>	<b>32</b>	136	24	<b>79</b>	<b>8</b>
	verb	201	<b>32</b>	<b>142</b>	<b>27</b>	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 Kolloquiums 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 International 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 Evaluation 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. Gładalski, 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 Processing*, RANLP, INCOMA Ltd., Shoumen, Bulgaria, 2023, pp. 641–647. URL: [http://dx.doi.org/10.26615/978-954-452-092-2\\_069](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, greco: 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. Shliachko, 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. Hadrow, 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:** 'nouthetéseis'  
**synonyms:** ['paramuthía', 'protropé', 'parakéleusis', 'parórmēsis', 'parokusmós', 'peithó', 'pístis', 'kéntron', 'múops', 'paráinēsis']

**word:** 'atimázō'  
**synonyms:** ['kataiskhúnō', 'aischúnō', 'atimóō', 'atimáo']

**word:** 'theosebēs'  
**synonyms:** ['deisidaímōn', 'eusebēés', 'eúphēmos', 'pístos']

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

**word:** 'trophé'  
**synonyms:** ['deípnōn', 'edódé', 'sítos', 'édesma']

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

**word:** 'iskhurós'  
**synonyms:** ['drastérios', 'karterós', 'energēs', 'rhōmaléos', 'krataíos', 'óbrimos', 'sthenarós', 'kraterós']

**word:** 'oknērōs'  
**synonyms:** ['phoberōs', 'deilōs']

## Declaration on Generative AI

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

# Author Index

- Acharjee, Prima, 206  
Alagni, Aurora, 4  
Albano, Paolo, 636  
Albertelli, Lisa Sophie, 12  
Alzetta, Chiara, 21  
Amaral, Patrícia, 561  
Amin, Muhammad Saad, 31  
Angioi, Alessandro, 570  
Anselma, Luca, 55, 924  
Arici, Nicola, 480  
Augello, Lorenzo, 12  
Auriemma, Serena, 83
- Bacco, Luca, 105, 911  
Bak, Joanna Ewa, 582  
Balducci, Gianmaria, 45  
Balestrucci, Pier Felice, 55, 924  
Ballatore, Andrea, 434  
Barba, Edoardo, 747  
Barbini, Matilde, 64  
Barone, Mariano, 995  
Barrón-Cedeño, Alberto, 263, 456  
Basile, Pierpaolo, 469, 796, 805  
Basile, Valerio, 31, 55, 384, 595, 603, 1028  
Basili, Roberto, 659, 775  
Bassi, Davide, 670  
Bassignana, Elisa, 815  
Bejgu, Andrei Stefan, 760  
Bellandi, Andrea, 595  
Bentivogli, Luisa, 860, 899  
Bernardi, Raffaella, 1037, 1121  
Biagetti, Erica, 647  
Bistarelli, Stefano, 314  
Block, Aleese, 363  
Bodlovic, Petar, 1158  
Bombieri, Marco, 72  
Bondielli, Alessandro, 83, 95, 423, 1121  
Bonetta, Giovanni, 1121  
Bonfigli, Agnese, 105, 911  
Bonomo, Tommaso, 760  
Bosco, Cristina, 1, 155  
Bottino, Andrea, 680  
Braga, Marco, 539  
Brasolin, Paolo, 595  
Bratières, Sébastien, 735  
Bressan, Veronica, 64, 1007  
Brigada Villa, Luca, 116, 647  
Broneske, David, 1018
- Brunato, Dominique, 105, 408, 837  
Brutti, Alessio, 860
- Calderaro, Silvio, 124  
Calvi, Giulia, 12  
Canale, Lorenzo, 134  
Capone, Luca, 143  
Cappello, Caterina Maria, 155  
Cariddi, Stefano, 625  
Carta, Salvatore Mario, 169  
Cascone, Lucia, 849  
Cassotti, Pierluigi, 177  
Castaldo, Antonio, 186  
Casula, Camilla, 194, 582  
Catalano, Luca, 206  
Ceccarelli, Cristian, 216  
Celli, Fabio, 225  
Cettolo, Mauro, 860  
Chesi, Cristiano, 64, 512, 1007, 1063  
Chierchiello, Elisa, 233  
Chiocchetti, Elena, 371  
Chiril, Patricia, 233  
Ciaccio, Cristiano, 245  
Cignarella, Alessandra Teresa, 603  
Ciletti, Michele, 256  
Ciminari, Debora, 263  
Cimmino, Martin, 636  
Clementelli, Annachiara, 12, 647  
Colombi, Anna Erminia, 292  
Colombi, Filippo, 273  
Colosi, Leonardo, 760  
Combei, Claudia Roberta, 282, 647  
Concas, Filippo, 169  
Conia, Simone, 747  
Corbetta, Claudia, 12, 292  
Cozzini, Marta, 304  
Creangă, Claudiu, 397  
Crispino, Filippo, 105  
Croce, Danilo, 659, 775  
Cuccarini, Marco, 314
- D'Alì, Sabrina, 155  
D'Angelo, Caterina, 327  
D'Asaro, Federico, 206, 680  
Da San Martino, Giovanni, 550  
Daffara, Agnese, 337  
Damiano, Rossana, 314  
Dampanaboina, Sai Teja, 352

De Cristofaro, Domenico, 363  
 De Luca, Ernesto William, 352, 526  
 Dell'Orletta, Felice, 105, 124, 245, 408, 870, 911  
 Delsanto, Matteo, 722  
 Di Bratto, Martina, 689  
 Di Fabio, Andrea, 595  
 Di Natale, Paolo, 371  
 Di Nuovo, Elisa, 155  
 Di Palma, Eliana, 384, 595  
 Dini, Luca, 408  
 Dinu, Anca, 397  
 Domenichelli, Lucia, 408  
 Draetta, Lia, 314  
 Duca, Giovanni, 1037  
  
 Enza, Messina, 45  
 Ergoli, Salvatore, 423  
 Ertaş, İrem, 849  
  
 Fallucchi, Francesca, 490  
 Farina, Andrea, 434  
 Farinetti, Laura, 134  
 Favalli, Andrea, 1148  
 Fedele, Domenico, 760  
 Fenu, Gianni, 169  
 Ferrando, Chiara, 886  
 Ferraresi, Adriano, 456  
 Fersini, Elisabetta, 45, 445, 625, 965, 974  
 Fiumanò, Beatrice, 314  
 Florescu, Andra-Maria, 397  
 Formichi, Guido, 1007  
 Fossat, Eugenio, 206  
 Frenda, Simona, 603  
 Fusco, Achille, 64, 512  
  
 Gabrieli, Giuliano, 384  
 Gaido, Marco, 860  
 Gajo, Paolo, 456  
 Galassi, Andrea, 747  
 Gallo, Simone, 550  
 Gamini, Sai Nishchal, 352  
 Gasparini, Francesca, 445  
 Gerevini, Alfonso Emilio, 480  
 Germani, Giovanni, 805  
 Ghiotto, Alessandro, 539  
 Ghizzota, Eleonora, 469  
 Giannone, Cristina, 1148  
 Ginevra, Riccardo, 647  
 Gioffré, Luca, 760  
 Giommarelli, Petra, 186  
 Giovannetti, Emiliano, 595  
  
 Giuliani, Alessandro, 169  
 Giulietti, Nicola, 206  
 Gjinika, Ejdis, 480  
 Gozzi, Manuel, 490  
 Grasso, Francesca, 500  
 Grazioso, Marco, 689  
 Gregori, Lorenzo, 955  
 Gretter, Roberto, 860  
 Guglielmetti, Mario, 155  
  
 Iaquinta, Tommaso, 512  
 Iurescia, Federica, 4, 12  
  
 Jagadeesh, Achal, 526  
 Jamshed, Minal, 206  
 Jarquín-Vásquez, Horacio Jesús, 31  
 Jezek, Elisabetta, 1, 603  
 Jiang, Chunyang, 826  
  
 Kasela, Pranav, 539  
 Khomyn, Olha, 983  
 Knez, Timotej, 785  
 Kołos, Anna, 582  
 Kunwar, Karishma, 352  
  
 Labruna, Tiziano, 550  
 Larson, Joseph E., 561  
 Lazzaroni, Ruggero Marino, 570  
 Lebani, Gianluca, 143, 710  
 Lenci, Alessandro, 83, 95, 143, 327, 423, 613, 933, 1121  
 Leonardelli, Elisa, 582  
 Levantesi, Marco, 352, 526  
 Lewinski, Marcin, 1158  
 Lilli, Livia, 735  
 Litta, Eleonora, 4, 595, 1093  
 Lo Giudice, Simona, 31  
 Lo, Soda Marem, 55, 603  
 Locci, Stefano, 500  
 Lombardi, Agnese, 613  
 Loreggia, Andrea, 480  
 Louf, Thomas, 582  
  
 Macagno, Alberto, 699  
 Madeddu, Marco, 636, 886  
 Madge, Chris, 1037  
 Maffia, Marta, 1055  
 Magazzù, Giuseppe, 625, 974  
 Magnini, Bernardo, 636, 747, 1121, 1158  
 Mambrini, Francesco, 595  
 Mammarella, Andrea Maurizio, 647

Manca, Marco Manolo, 169  
 Mancini, Azzurra, 689  
 Marchesi, Beatrice, 647  
 Marchi, Simone, 595  
 Margiotta, Daniele, 659  
 Marino, Erik Bran, 670  
 Marras, Mirko, 169  
 Marras, Roberto, 570  
 Mastellari, Virginia, 647  
 Matassoni, Marco, 860  
 Mauro, Valeria, 689  
 Mazzei, Alessandro, 55, 924  
 Mazzei, Samuele, 699  
 Mazzoli, Simone, 710  
 McGillivray, Barbara, 434  
 Mensa, Enrico, 722  
 Merlo, Paola, 826  
 Merone, Mario, 105, 911  
 Messina, Alberto, 134  
 Miaschi, Alessio, 124, 245, 1121  
 Michielon, Edoardo, 625  
 Micluța-Câmpeanu, Marius, 397  
 Mihail, Andreiana, 397  
 Milani, Tommaso, 722  
 Miliani, Martina, 83  
 Miranda, Michele, 735  
 Montemagni, Simonetta, 21  
 Monti, Johanna, 186  
 Moretti, Giovanni, 292, 595, 1093  
 Moroni, Luca, 747, 760  
 Moscato, Emanuele, 815  
 Moscato, Vincenzo, 995  
 Mousavi, Seyed Mahed, 1028  
 Mousavian Anaraki, Seyed Alireza, 775  
 Muhammad, Iftikhar, 785  
 Mura, Piergiorgio, 169  
 Muratore, Elisa, 582  
 Musacchio, Elio, 796, 805  
 Muscianisi, Domenico Giuseppe, 1093  
 Muti, Arianna, 815  
 Márquez Villacís, Juan José, 680  
  
 Nastase, Vivi, 826  
 Navigli, Roberto, 747, 760  
 Nawar, Mohamed Nabih Ali Mohamed, 860  
 Negri, Matteo, 860, 899  
 Neri, Sofia, 64, 1007  
 Nozza, Debora, 815  
  
 Oliverio, Michael, 55, 924  
 Orsini, Massimiliano, 837  
  
 Özkan, Aylin, 849  
  
 Pagano, Adriana, 233  
 Paglione, Luca, 83  
 Palmero Aprosio, Alessio, 699  
 Pantaleo, Michele, 206  
 Papalia, Giuseppe Francesco, 105  
 Papalia, Rocco, 105  
 Papi, Sara, 860  
 Pappacoda, Gianmarco, 747  
 Papucci, Michele, 870  
 Pasini, Sveva Silvia, 886  
 Pasqualini, Marco, 625  
 Passaro, Lucia C., 83, 95, 933, 1121  
 Passarotti, Marco, 292, 595  
 Patarnello, Stefano, 735  
 Patel, Sahithya Narayanaswamy, 526  
 Patti, Viviana, 31, 55, 603, 636, 886  
 Pecchia, Leandro, 105, 911  
 Piccini Bianchessi, Maria Letizia, 64  
 Piergentili, Andrea, 899  
 Pilzer, Andrea, 539  
 Piperno, Ruben, 105, 911  
 Pirrone, Roberto, 1074  
 Pisano, Simone, 169  
 Poesio, Massimo, 1037  
 Polignano, Marco, 1, 352  
 Polizzi, Daniele, 456  
 Ponzetto, Simone Paolo, 72  
 Porta, Stefano Vittorio, 924  
 Postiglione, Marco, 995  
 Pouplier, Marianne, 363  
 Presutti, Valentina, 314  
 Proietti, Mattia, 933  
 Pulerà, Francesca, 625, 974  
 Puliga, Michelangelo, 570  
 Putelli, Luca, 480  
  
 Radicioni, Daniele P., 722  
 Raganato, Alessandro, 216, 539  
 Ramponi, Alan, 337  
 Ranaldi, Federico, 942  
 Ranaldi, Leonardo, 942  
 Ravelli, Andrea Amelio, 955  
 Redaelli, Arianna, 1084  
 Resta, Michele, 636  
 Riccardi, Giuseppe, 983  
 Riccio, Giuseppe, 995  
 Rigutini, Leonardo, 1171  
 Rinaldi, Matteo, 955  
 Rizzi, Giulia, 445, 625, 965, 974

Rizzo, Giuseppe, 206, 680  
Roccabruna, Gabriel, 983  
Romagnoli, Raniero, 1148  
Romano, Antonio, 995  
Rospocher, Marco, 72, 785  
Rossi, Sarah, 64, 1007  
Rovera, Marco, 225  
Russo, Fabrizio, 105  
Russo, Valentina, 689  
Ruzzetti, Elena Sofia, 1148

Safikhani, Parisa, 1018  
Saggion, Horacio, 304  
Saibene, Aurora, 445  
Sajedinia, Maryam, 1028  
Samo, Giuseppe, 826  
Sanguinetti, Manuela, 1  
Sanna, Davide, 570  
Sansonetti, Franco, 31  
Sarti, Gabriele, 245  
Sarzotti, Marika, 1037  
Savoldi, Beatrice, 899  
Scalena, Daniel, 625, 965, 974  
Schettino, Loredana, 1048, 1055  
Sciolette, Flavia, 595  
Scirè, Alessandro, 760  
Scotta, Stefano, 134  
Scozzaro, Calogero Jerik, 722  
Semeraro, Giovanni, 352, 469, 526, 796, 805  
Serina, Ivan, 480  
Sgrizzi, Tommaso, 64, 1007, 1063  
Shankar, Chinmayi Ravi, 526  
Siciliani, Lucia, 469, 796, 805  
Siragusa, Irene, 1074  
Sormani, Alberto, 625, 974  
Sprugnoli, Rachele, 1084, 1093  
Stamile, Claudio, 625  
Stemle, Egon Waldemar, 371  
Stranisci, Marco Antonio, 603  
Strapparava, Carlo, 273  
Suozzi, Alice, 143, 710

Taddei, Andrea, 327

Tahmasebi, Nina, 177  
Tamburini, Fabio, 1102, 1112  
Tealdo, Gabriele, 699  
Testa, Davide, 1121  
Tonelli, Sara, 194, 337, 582  
Torrioni, Paolo, 747  
Troncone, Luisa, 1135  
Tăbușcă, Ștefana Arina, 397  
Ubaldi, Asia Beatrice, 805  
Vadalà, Gianluca, 105  
Vardanega, Agnese, 384  
Varvara, Rossella, 955  
Vassallo, Marco, 384  
Vecellio Salto, Sebastiano, 582  
Venturi, Giulia, 870  
Vieira, Renata, 670  
Vietti, Alessandro, 363, 1048  
Vignoli, Anna, 282  
Vitale, Vincenzo Norman, 1048, 1055  
Viviani, Marco, 216, 539

Xompero, Giancarlo A., 1148  
Xu, Lu, 760

Yavuz, Mehmet Can, 849  
Yin, Michele, 983

Zambotto, Lorenzo, 699  
Zampetta, Silvia, 647  
Zanchi, Chiara, 647, 886  
Zane, Lorenzo, 722  
Zaninello, Andrea, 1158  
Zanoli, Roberto, 636, 747  
Zanollo, Asya, 512, 1007, 1063  
Zanzotto, Fabio Massimo, 1148  
Zappulla, Francesco, 282  
Zeinalipour, Kamyar, 512  
Zugarini, Andrea, 1171  
Žitnik, Slavko, 785