

<https://doi.org/10.17323/jle.2026.27552>

Category-Dependent Effectiveness of Web-Based and AI-Generated Usage Examples in Modern Greek Pedagogical Lexicography

Lamprini Lourida ¹, Katerina Alexandri ²

¹ University of Patras, Patras, Greece

² Democritus University of Thrace, Alexandroupolis, Greece

ABSTRACT

Background. Usage examples are an important feature of pedagogical dictionaries, as they help learners understand meaning, context, and lexical use. Although web-based resources and large language models offer new possibilities for generating dictionary examples, their pedagogical suitability has not yet been sufficiently examined in Modern Greek.

Purpose. To address this gap, this exploratory study compares Web-based and AI-generated usage examples across three lexically challenging categories in Modern Greek: polysemous verbs and nouns, and idiomatic expressions. It examines whether and how the pedagogical suitability of each source varies according to lexical complexity for learners aged 11-15 and explores their potential integration in pedagogical lexicography.

Materials and Methods. A small-scale exploratory dataset of 81 usage examples was collected from Web sources and generated by GPT-5.3 Instant and DeepSeek-V3 and was evaluated using a combined framework drawing on GDEX criteria and pedagogical appropriateness indicators on a three-point ordinal scale. A learner validation phase with 41 students (aged 11-15) was also conducted.

Results. Within the analysed dataset, the findings suggest category-related differences across lexical categories. Web-derived examples showed greater semantic diversity and higher correct meaning identification rates in the learner evaluation (56.50%), particularly for figurative and idiomatic uses. AI-generated examples, especially those produced by GPT-5.3, showed higher intelligibility scores and performed more consistently for semantically stable items such as polysemous nouns, but exhibited limitations in idiomatic language, with correct meaning identification for GPT-5.3 dropping to 32.37% for specific idiomatic items. DeepSeek-generated examples showed the most pronounced limitations within the analysed categories, particularly in typicality and pedagogical appropriateness.

Conclusion. The findings suggest that, within the analysed dataset, the pedagogical value of usage examples depends not only on their source, but also on the type of lexical complexity involved. Neither Web-based nor AI-generated examples can independently meet the requirements of a learner-oriented dictionary. Instead, the findings support a hybrid category-sensitive approach that combines Web-based authenticity, expert curation, selective AI assistance, and attention to learners' actual comprehension. The methodological framework developed here may offer a basis for future comparative work in other morphologically complex and idiomatically rich languages, though transferability remains to be tested.

KEYWORDS

pedagogical lexicography, AI-generated examples, learner dictionaries, hybrid lexicographic design, under-resourced languages

Citation: Lourida, L., & Alexandri, K. (2026). Category-dependent effectiveness of web-based and AI-generated usage examples in modern Greek pedagogical lexicography. *Journal of Language and Education*, 12(1), 160-174. <https://doi.org/10.17323/jle.2026.27552>

Correspondence:
Lamprini Lourida
l.lourida@ac.upatras.gr

Received: July 03, 2025

Accepted: March 16, 2026

Published: March 31, 2026

INTRODUCTION

Lexicography plays a central role in language learning and language descrip-

tion, particularly in educational contexts, where dictionaries function as tools for vocabulary development and language awareness. Traditionally, lexicograph-



ic resources have relied on manually collected corpus data and expert-generated usage examples. However, the rapid expansion of digital communication and the increasing availability of computational tools have significantly transformed lexicographic practices. In particular, the emergence of large language models (LLMs) and the use of the Web as a linguistic resource have introduced new possibilities for generating usage examples and updating dictionary content more dynamically (Lew, 2024; de Schryver, 2024; Rundell, 2024).

These developments are especially relevant for pedagogical lexicography, where usage examples play a crucial role in supporting learners' understanding of meaning, context, and syntactic behavior (Atkins & Rundell, 2008; Lew, 2015). High-quality examples help learners interpret polysemous words, recognize figurative meanings, and understand idiomatic expressions. At the same time, generating suitable examples remains a challenging and time-consuming task, particularly for morphologically rich languages and educational contexts that require age-appropriate and pedagogically meaningful content.

Modern Greek represents a particularly relevant case in this respect. Despite a long lexicographic tradition and the development of several pedagogical dictionaries in recent decades, there remains a limited availability of up-to-date, learner-oriented digital dictionaries designed specifically for school-aged users. Earlier efforts to develop pedagogical dictionaries for Modern Greek have contributed significantly to educational lexicography (e.g., Vakalopoulou, 2000; Vakalopoulou & Iordanidou, 2001; Iordanidou et al., 2007; Gavrilidou et al., 2007; Kapsalis et al., 2007), yet many of these resources are no longer aligned with contemporary digital standards or current language use. As language evolves rapidly in digital environments, pedagogical dictionaries require continuous updating and adaptation to reflect authentic, contemporary usage. Recent studies in Greek lexicography have begun to explore how AI-assisted methods are relevant for Modern Greek lexicographic practice and can support the resolution of language-related questions (Alexandri, 2025; Alexandri & Iordanidou, 2025).

The Web has increasingly been used as a large, dynamic corpus that provides access to authentic language data across a wide range of contexts (Kilgarriff, 2007; Davies, 2018; Hoenen et al., 2020; Gatto, 2025). Web-based examples can capture emerging meanings, contemporary usage patterns, and stylistic variation that may not be present in traditional corpora, including variation across registers and domains (Davies, 2018; Arkhangelskiy, 2019; Laippala et al., 2021). At the same time, Web-derived data require careful selection and pedagogical adaptation, as authenticity alone does not guarantee suitability for learners (Kilgarriff, 2007; Fuertes-Olivera, 2012).

Alongside Web-based approaches, recent studies have explored the potential of large language models for lexicographic tasks, including definition writing and example generation (Lew, 2023; Phoodai & Rikk, 2023; Jakubiček & Rundell, 2023; Rundell, 2024; Kosem et al., 2024). Previous research suggests that LLMs can produce grammatically well-formed dictionary example sentences (Cai et al., 2024; Almeman et al., 2024; Lewandowska-Tomaszczyk & Pawłowski, 2025), but concerns remain regarding semantic accuracy, naturalness, and pedagogical appropriateness, particularly in cases involving polysemy and idiomatic language (Merx et al., 2024; Durward & Thomson, 2024).

Research in this area so far has tended mainly to examine web-based usage examples and LLM-generated examples separately or to evaluate LLM outputs against traditional lexicographic resources (e.g., Kilgarriff & Grefenstette, 2003; Fuertes-Olivera, 2012; Crosthwaite & Baisa, 2023; Almeman et al., 2024; Nasution & Onan, 2024). Crucially, existing research has not sufficiently examined how the usefulness of each source changes depending on lexical complexity and learner comprehension. Systematic, pedagogically orientated comparisons between Web-based and AI-generated usage examples remain limited, particularly for morphologically rich and moderately under-resourced languages such as Modern Greek (Krstev & Stanković, 2023). The present study aims to address these gaps in the context of Modern Greek pedagogical lexicography by comparing and evaluating the pedagogical suitability of usage examples sourced from Web resources and large language models. Rather than evaluating grammatical correctness alone, the study focuses on pedagogical usability for learners aged 11-15, with particular attention to polysemy, figurative meaning, and idiomatic expressions. This focus allows for a more nuanced evaluation of example quality in educational lexicographic contexts.

The study addresses the following research questions:

- RQ1:** How do Web-based and AI-generated usage examples compare in terms of their pedagogical suitability for learners aged 11-15, across different categories of lexical complexity?
- RQ2:** What patterns of variation emerge across lexical categories, and what implications do these patterns have for the integration of different data sources in pedagogical lexicographic practice?

By proposing a category-sensitive approach to evaluating usage examples in Modern Greek pedagogical lexicography, the study aims to contribute to ongoing discussions on the role of Web-based and AI-generated data in lexicographic practice. Rather than assuming that one source is uniformly more effective than another, the study explores how different lexical categories interact with different types of linguistic data. While the analysis focuses on Modern Greek, the

methodological framework and findings may provide a basis for future comparative research on pedagogical lexicography in other morphologically rich and moderately under-resourced languages.

Based on previous research, three hypotheses were formulated. First, it was hypothesized that Web-based examples would demonstrate greater semantic variability and contextual richness than AI-generated examples, particularly in cases involving figurative and idiomatic language, where authentic data are expected to better reflect the range of attested usage (Kilgarriff, 2007; Fuertes-Olivera, 2012). Second, it was hypothesized that AI-generated examples would show higher levels of intelligibility and structural consistency, given the tendency of LLMs to produce well-formed, prototypical sentences (Cai et al., 2024; Almeman et al., 2024). Third, it was hypothesized that the relative performance of each source would not be uniform across lexical categories, but would vary systematically depending on the type of lexical complexity involved, with AI-generated examples expected to perform less reliably in cases of polysemy and idiomaticity compared to semantically stable items (Merx et al., 2024; Arnett & Bergen, 2025). Testing these hypotheses is theoretically meaningful not because their direction is unexpected, but because the study provides the initial exploratory evidence of how these tendencies vary across polysemous verbs, polysemous nouns, and idiomatic expressions in a learner-centred pedagogical lexicographic context.

LITERATURE REVIEW

Usage Examples in Pedagogical Dictionaries

Pedagogical dictionaries are designed to support language learning by addressing the needs of specific user groups, particularly young learners and second language users. Historically, the term “pedagogical dictionary” referred primarily to monolingual dictionaries developed for learners of English as a foreign language. Today, the concept has broadened to include dictionaries designed for school-aged users, including native speakers in primary and early secondary education (Verlinde & Binon, 2009; Chi, 2022). Pedagogical lexicography therefore operates at the intersection of theoretical linguistics, applied linguistics, language education, and lexicographic practice.

Pedagogical dictionaries differ from general-purpose dictionaries both at the macrostructural and microstructural levels. At the macrostructural level, entries are selected and organized according to learners’ needs and curricular relevance. At the microstructural level, definitions are simplified, grammatical information is selectively presented, and usage examples play a central role in supporting comprehension and language development (Atkins & Rundell, 2008; Verlinde & Binon, 2009). In this context, usage examples function not

merely as illustrative additions but as essential components that facilitate learners’ understanding of meaning, contextual variation, and syntactic behaviour.

The selection and formulation of usage examples have long been recognized as a critical challenge in pedagogical lexicography. Traditionally, lexicographers have relied on corpus-based approaches, extracting authentic sentences from language corpora and adapting them to meet pedagogical requirements (Tarp & Gouws, 2020). However, raw corpus examples often require editing to ensure clarity, relevance, and suitability for learners. This need for systematic selection led to the development of evaluation frameworks such as GDEX (Good Dictionary Examples), which provide criteria for identifying high-quality usage examples (Kilgarriff et al., 2008).

According to the GDEX framework, high-quality usage examples should meet several core criteria. These include typicality, referring to the representation of common and representative usage patterns; informativeness, which ensures that the example helps users understand meaning and usage; and intelligibility, which emphasizes readability and avoidance of complex structures that may hinder comprehension (Kilgarriff et al., 2008). These criteria aim to ensure that usage examples serve both lexicographic and pedagogical functions.

Subsequent research has proposed additional criteria to further refine the evaluation of usage examples. Kosem et al. (2019), for instance, introduced naturalness, authenticity, and self-containment as complementary dimensions. “Naturalness” refers to examples that reflect conventional language use; “authenticity” emphasizes the importance of real-world data; and self-containment highlights the need for examples that are understandable without additional context. Together, these criteria contribute to a more comprehensive framework for evaluating usage examples in learner-orientated dictionaries.

Despite these developments, selecting pedagogically appropriate examples remains challenging. Authentic examples may be linguistically complex, while simplified examples may lack naturalness or contextual richness. Introducing new sources of linguistic data, such as Web-based corpora and large language models, into lexicographic practice makes these tensions particularly relevant. As a result, evaluating usage examples requires balancing authenticity, clarity, and pedagogical suitability. This study builds on existing evaluation frameworks for usage examples by applying GDEX-based criteria (Kilgarriff et al., 2008; Kosem et al., 2019) alongside pedagogical considerations tailored to school-aged learners (Atkins & Rundell, 2008). Critically, the GDEX framework alone is insufficient for the present study, as it was not designed with young learner populations in mind; the addition of pedagogical appropriateness as a criterion is therefore theoretically necessary to address the specific cognitive and linguistic needs of learners aged 11-15.

The Web as a Corpus for the Collection of Usage Examples

The rapid expansion of digital communication has transformed the Web into one of the largest and most dynamic sources of linguistic data available today. Unlike traditional corpora, which are often curated and periodically updated, the Web provides continuously evolving language use across diverse communicative contexts, including news articles, blogs, fora, and social media platforms. This diversity enables researchers and lexicographers to access authentic and contemporary language use, making the Web a valuable resource for lexicographic practice (Kilgarriff & Grefenstette, 2003; Kilgarriff, 2007; Gatto, 2025).

The increasing availability of Web-based data has contributed to a shift in lexicographic methodology, particularly in pedagogical lexicography. Corpus-based approaches have already transformed dictionary compilation by enabling lexicographers to analyse systematic patterns of language use and develop user-orientated resources grounded in authentic linguistic data (Davies, 2018; Arkhangelskiy, 2019). The Web extends this potential further by offering access to emerging linguistic patterns, recent lexical developments, and informal language varieties that are often under-represented in traditional corpora. This development aligns with a broader movement within electronic lexicography toward dynamic, usage-based, and user-orientated resources that integrate large-scale, heterogeneous linguistic data (Tarp, 2014; Tarp & Gouws, 2020; Lew, 2024; Rundell et al., 2025).

Authentic usage examples derived from real-world contexts are widely recognized as essential components of pedagogical dictionaries (Tarp, 2014). Such examples help learners understand how words function in natural communicative settings and support vocabulary development by illustrating syntactic patterns and contextual variation. Web-based data, in particular, provide opportunities to capture contemporary usage and stylistic variation across different registers. As Fuertes-Olivera (2012) notes, the accessibility and breadth of Web-based resources allow lexicographers to retrieve examples reflecting current language use, including informal and conversational language that may not appear in traditional corpora, making the Web a valuable resource for lexicographic practice that captures contemporary usage and stylistic variation across registers (Kilgarriff & Grefenstette, 2003; Davies, 2018; Gatto, 2025).

Despite these advantages, the use of the Web as a linguistic resource also presents methodological challenges. Search engine results are inherently unstable and may vary depending on search operators, ranking algorithms, and temporal factors, which complicates reproducibility and frequency analysis (Kilgarriff, 2007). In addition, Web-based retrieval often yields heterogeneous results, including dictionary entries,

duplicated content, and contextually incomplete sentences that require manual filtering. These challenges are further amplified in morphologically rich languages such as Modern Greek, where retrieving representative examples may require searches across multiple inflected forms.

These limitations highlight the need for careful curation and evaluation when using Web-derived examples in pedagogical lexicography. While the Web offers authenticity and linguistic diversity, selecting pedagogically appropriate examples remains a labour-intensive process requiring expert judgement. This need for expert filtering and adaptation is directly relevant to the design of the present study: Web-based retrieval was treated not as a neutral or fully automatic process, but as a first stage requiring systematic pedagogical evaluation. The challenge of manual filtering and the instability of search engine results thus directly motivated the methodological choices described in the following section.

AI-Generated Usage Examples in Lexicographic Practice

Alongside Web-based approaches, recent studies have explored the potential of large language models for lexicographic tasks, including definition writing and example generation (Lew, 2023; Phoodai & Rikk, 2023; Jakubiček & Rundell, 2023; Rundell, 2024; Kosem et al., 2024; Garipov et al., 2026; Ide et al., 2026). Previous research suggests that LLMs can produce grammatically well-formed dictionary example sentences (Cai et al., 2024; Almeman et al., 2024; Lewandowska-Tomaszczyk & Pawłowski, 2025), but concerns remain regarding semantic accuracy, naturalness, and pedagogical appropriateness, particularly in cases involving polysemy and idiomatic language (Merx et al., 2024; Durward & Thomson, 2024).

Studies on LLM performance in morphologically complex and less-resourced languages further suggest that model outputs may be less reliable in these contexts, including in the handling of idiomatic expressions and figurative language (Arnett & Bergen, 2025; Abdelrahman, 2024; Beliga & Filipović Petrović, 2024; Marković & Stanković, 2025). These findings are particularly relevant for Modern Greek, which combines a rich morphological system with an extensive idiomatic repertoire and is less well-represented in the training corpora of most major LLMs (Alexandri & Iordanidou, 2025).

Taken together, this body of research justifies evaluating AI-generated examples not only for grammaticality but also for semantic precision, idiomaticity, and learner comprehension. It also underscores why a zero-shot prompting strategy is appropriate for approximating typical user interaction conditions, while simultaneously making clear that such a design does not capture the full potential of LLMs under optimized or expert-guided prompting configurations.

METHOD

Data Collection from Web Sources

To collect authentic usage examples, a structured Web-based retrieval procedure was implemented using the Google search engine. The Web was treated as a large, dynamic corpus of contemporary language use, providing access to a wide range of communicative contexts. To enhance retrieval precision and ensure consistency, predefined search strings were constructed for each lexical item. These included combinations of common inflected forms using Boolean operators (e.g., «θερίζει» OR «θέρισε» OR «θερίζουν» OR «θέρισαν»), as well as quotation marks for phrase-based searches. This approach was adopted to address the morphological complexity of Modern Greek and to increase the likelihood of retrieving representative usage examples. Searches were restricted to Greek-language content published between 2015 and 2026 to capture contemporary language use while minimizing outdated or less relevant examples. In addition, retrieval was limited to publicly accessible online sources, primarily journalistic and informational websites, which are known to reflect current usage patterns. The retrieval process followed a three-stage procedure:

- (1) Initial search using predefined search strings
- (2) Screening of results for relevance and contextual completeness
- (3) Selection of examples based on pedagogical suitability

For each lexical item, the first three unique and contextually complete results were selected for analysis. In cases where duplicate or near-identical examples were identified (e.g., syndicated news content), these were excluded and replaced by the next available unique result. Examples were also excluded if they lacked sufficient contextual information, contained highly specialized or technical vocabulary, or required extensive background knowledge beyond the target learner level. The final dataset consisted of 81 manually curated and annotated usage examples across all lexical categories. This procedure resulted in a curated set of Web-based examples that reflect authentic and contemporary usage while maintaining relevance for learners aged 11-15. However, several important methodological limitations should be noted. The reliance on a general-purpose search engine introduces variability related to ranking algorithms, temporal factors, and personalization, all of which affect reproducibility. The decision to select the first three relevant results per query may also introduce a selection bias that privileges certain registers or discourse types, such as journalistic and informational language, over others. These factors may shape the observed patterns in ways that are difficult to fully control, and this should be taken into account when interpreting the results.

AI-Generated Examples

To evaluate the potential of large language models (LLMs) in pedagogical lexicography, usage examples were generated using two widely accessible systems: OpenAI's GPT-5.3 Instant (web interface) and DeepSeek-V3. These models were selected due to their increasing use in language-related tasks and their relevance in recent lexicographic research (Lew, 2023; Rundell, 2024; Kosem et al., 2024; Zhong, 2025). All examples were generated during a single data collection phase (March 2026) using the web-based interfaces of both systems. Default system settings were applied in both cases, and no additional parameter tuning (e.g. temperature adjustments) or system-level instructions were used. This approach was adopted to ensure consistency across models and to approximate typical end-user interaction conditions. A zero-shot prompting strategy was employed to assess baseline model performance without task-specific training or prompt engineering. This choice was intended to examine how publicly accessible LLMs perform under realistic, minimally mediated user conditions, rather than under optimized experimental prompting. More specifically, a Zero-Shot General + Audience (ZS-GEN+AUD) prompting approach was used. All prompts were formulated in Greek and explicitly specified:

- (1) the role of the system as a pedagogical dictionary;
- (2) the target language (Modern Greek);
- (3) the intended audience (learners aged 11-15);
- (4) the task (generation of usage examples for a given lexical item).

The prompt used for all lexical items followed a consistent structure, with variation only in the target word: «Παράγαγε 3 παραδείγματα χρήσης στα νέα ελληνικά για τη λέξη "X". Τα παραδείγματα πρέπει να απευθύνονται σε μαθητές/τριες ηλικίας 11-15 ετών.» (Produce three usage examples in Modern Greek for the word "X". The examples should be suitable for students aged 11-15). Each lexical item was prompted separately in both models. For each prompt, three usage examples were generated. No iterative prompting, regeneration, or few-shot examples were used to maintain comparability across models and to evaluate baseline performance under controlled conditions.

Several important design constraints should be acknowledged. First, the study evaluates only one prompting configuration and cannot be used to draw conclusions about the full potential of LLMs under optimized prompting, few-shot prompting, regeneration, or expert-guided prompt engineering. Second, each lexical item was generated only once per model, which introduces output instability that is not controlled for; given that LLM outputs can vary across runs, this raises questions about the replicability of individual results. The findings should therefore be understood as reflecting LLM performance under this specific prompting configura-

tion and these specific conditions, rather than as general assessments of LLM capability in lexicographic tasks.

All generated outputs were subsequently screened and selected for analysis using the same inclusion criteria applied to Web-based examples, namely clarity, completeness, and relevance for the target learner group. This ensured methodological consistency across data sources while allowing for a direct comparison between Web-derived and AI-generated usage examples. This design enables evaluation of baseline model performance under realistic user conditions, while results should not be generalized beyond this specific prompting configuration.

Selection of Lexical Items

The selection of lexical items followed a purposive and criterion-based sampling strategy, aiming to capture representative cases of lexical complexity relevant to pedagogical lexicography in Modern Greek. The items were selected from polysemy-focused exercises in Modern Greek school textbooks (upper primary and lower secondary levels), where students are required to interpret multiple meanings and distinguish between literal, figurative, and idiomatic uses. This approach ensures that the selected items reflect documented areas of learner difficulty within the curriculum. The dataset includes lexical items across three categories of lexical complexity:

- (1) polysemous verbs (e.g., *θερίζω* [θerízo] and comparable items);
- (2) polysemous nouns with broad semantic range (e.g., *κρίση* [krísi] and similar cases);
- (3) idiomatic expressions derived from high-frequency verbs (e.g., *τρύω* [trío] and related expressions).

This categorization allows for a systematic comparison of how different types of lexical difficulty (polysemy, semantic variation across domains, and idiomaticity) affect the pedagogical suitability of Web-based and AI-generated usage examples. The selection of lexical items was guided by the following criteria:

- (1) documented difficulty in school-based language tasks;
- (2) frequency and relevance in everyday language use;
- (3) presence of multiple meanings or figurative extensions;
- (4) pedagogical relevance for learners aged 11-15.

Given the scope of the study, the selected items constitute analytical case studies rather than a representative sample of Modern Greek vocabulary. The dataset was designed to illuminate how the type of lexical complexity affects source suitability, not to characterize LLM or Web performance in general. Claims derived from the analysis therefore apply specifically to this set of items and should not be generalized to the wider Modern Greek lexicon.

Evaluation Framework and Annotation Procedure

To evaluate the pedagogical suitability of usage examples, a structured evaluation framework was developed, combining GDEX-based criteria with pedagogical considerations relevant to learners aged 11-15. Four evaluation criteria were applied: typicality, informativeness, intelligibility, and pedagogical appropriateness. Each criterion was operationalized using a three-point ordinal scale (0-2) to capture gradience in lexicographic and pedagogical quality. This approach was preferred over binary judgments, as lexicographic suitability is typically distributed along a continuum rather than falling into discrete categories.

The criteria were defined as follows:

- (1) Typicality referred to whether the example reflected common and natural usage in contemporary Modern Greek: 0 = atypical or unnatural usage, 1 = partially typical or somewhat marked usage, 2 = fully typical and representative usage.
- (2) Informativeness referred to the extent to which the example clearly illustrated the meaning of the lexical item: 0 = meaning unclear or poorly illustrated, 1 = meaning partially inferable but context limited, 2 = meaning clearly illustrated with sufficient context.
- (3) Intelligibility referred to the readability and linguistic accessibility of the example for learners aged 11-15: 0 = difficult to understand due to vocabulary or structure, 1 = generally understandable but somewhat complex, 2 = clear and easily understandable.
- (4) Pedagogical appropriateness referred to the suitability of the example for educational use: 0 = unsuitable for pedagogical use, 1 = usable with adaptation or teacher mediation, 2 = directly suitable for pedagogical dictionary use.

Each example was independently evaluated by two expert annotators with experience in Modern Greek lexicography and language education. Annotators assigned a score for each criterion and provided brief qualitative comments, particularly in borderline or ambiguous cases. Following independent evaluation, discrepancies between annotators were identified and discussed. Consensus was reached through collaborative review, guided by the operational definitions of each criterion. This two-stage procedure was designed to enhance consistency and reduce individual bias. Preliminary agreement between annotators was high across most criteria, with disagreements primarily occurring in cases involving idiomatic or context-dependent interpretation. For analytical purposes, total scores were calculated by summing the values across the four criteria (maximum score: 8). Based on the total score, examples were categorized as follows: 0-3: low pedagogical suitability, 4-6: moderate pedagogical suitability, 7-8: high pedagogical suitability.

This scoring system enabled a more systematic comparison of Web-based and AI-generated examples while preserving qualitative distinctions in pedagogical and lexicographic suitability.

Inter-rater reliability was assessed using weighted Cohen's kappa, appropriate for ordinal rating scales. Agreement was high across all evaluation criteria (weighted Cohen's $\kappa = .84-.91$), indicating very high to almost perfect agreement among the annotators. Disagreements occurred primarily in borderline instances involving idiomatic or context-dependent interpretation. These were resolved through structured discussion guided by the operational definitions of each criterion, in line with practice in small-scale qualitative lexicographic research (Kilgarriff et al., 2008; Kosem et al., 2019).

Learner Validation Questionnaire

A follow-up learner-centered evaluation was conducted with 41 lower secondary school students, aged 11-15. Participation was voluntary and conducted in accordance with ethical guidelines. Students were recruited from a single school in Peloponnese, Greece; they represented a homogeneous group in terms of age range and curriculum level, which limits the representativeness of the sample. For each lexical item, students were presented with three usage examples, one from each source (Web, GPT-5.3 and DeepSeek-V3). Where available, examples previously evaluated as highly pedagogically suitable (scores 7-8) were selected. In cases where no highly suitable AI-generated example was available for a given lexical item, the best available example from each source was included, to ensure comparability across sources. Students were asked to evaluate the usage examples using Likert-scale items in terms of clarity, meaning comprehension, and usefulness for understanding word usage. Responses were collected and analyzed using descriptive statistics to compare learner preferences across sources. No inferential statistical testing was conducted due to sample size constraints; differences between sources should therefore be interpreted as observed tendencies rather than statistically significant effects. Items for which complete response data were unavailable were excluded from the corresponding category-level analyses. This procedure aimed to complement the expert-based evaluation with learner-based evidence of the pedagogical effectiveness of usage examples.

Table 1

Mean Evaluation Scores for Polysemous Verbs

Source	Typicality	Informativeness	Intelligibility	Pedagogical Appropriateness	Mean Total
Web	1.9	1.9	1.7	1.6	7.1
GPT-5.3	1.5	1.6	1.9	1.5	6.5
DeepSeek-V3	1.2	1.4	1.8	1.2	5.6

Note. *θερίζω* [θerízo], *κλίνω* [klíno], *καίω* [kéo]

RESULTS

Evaluation of polysemous verbs *θερίζω* [θerízo], *κλίνω* [klíno], *καίω* [kéo]

The analysis of polysemous verbs suggested differences within the analysed dataset between Web-based and AI-generated examples in terms of semantic coverage and pedagogical suitability. Web-based examples received high scores for typicality and semantic breadth, particularly in figurative extensions. In the case of *θερίζω*, Web data captured metaphorical uses related to death, illness, and large-scale impact, reflecting contemporary journalistic discourse. For example: «*Θερίζει*» η γρίπη με εξαντλητικά συμπτώματα - εκατοντάδες παιδιά στα επείγοντα (The flu is sweeping through the population, causing severe symptoms - Hundreds of children in Emergency Rooms.). Similarly, *κλίνω* was predominantly attested in figurative meanings (e.g. tendency, decision-making), as in: *Οι ΗΠΑ «κλίνουν» προς το σενάριο πολέμου με το Ιράν* (The US is leaning toward a war scenario with Iran.). The verb *καίω* displayed both literal and extended meanings, including economic burden: *Οι λογαριασμοί του ρεύματος «καίνε» τα νοικοκυριά* (Electricity bills are placing a heavy financial burden on households.). In contrast, AI-generated examples showed systematic tendencies toward literalization and semantic restriction. GPT-5.3 produced generally well-formed and intelligible examples, but with uneven semantic coverage and occasional collocational errors, such as: *Η καταιγίδα θέρισε τις καλλιέργειες* (The storm destroyed the crops). Although the intended meaning ("destroyed") is inferable, the collocation is non-typical in Modern Greek. DeepSeek-V3 showed greater instability, including non-standard metaphorical extensions, as in: *Η ομάδα μπάσκετ θέρισε όλους τους αντιπάλους* (Basketball team defeated all its opponents.). Such examples were evaluated as non-typical in Modern Greek and received lower scores for typicality and pedagogical appropriateness. As shown in Table 1, these patterns are observed within the analysed sample and should be interpreted as tendencies rather than generalizable findings. Web-based examples achieved the highest scores across all criteria, particularly in typicality and informativeness. GPT-5.3 performed comparatively well in intelligibility but showed limitations in semantic precision. DeepSeek-V3 received the lowest scores, particularly in pedagogical appropriateness.

Evaluation of polysemous nouns κρίση [krísi], στάση [stási], γλώσσα [glósa]

For polysemous nouns, Web-based examples again demonstrated broad semantic coverage and contextual richness. The noun *κρίση* illustrated multiple domains (economic, psychological, institutional), as in: *Η οικονομική κρίση επηρέασε πολλές οικογένειες* (The economic crisis affected many families). At the same time, evaluative meanings were also attested: *Η κρίση του δασκάλου για την εργασία μας ήταν θετική* (The teacher's assessment of our work was positive). Similarly, *στάση* and *γλώσσα* showed both literal and abstract meanings across different registers: *Περιμέναμε στη στάση του λεωφορείου* (We waited at the bus stop). / *Η στάση του απέναντι στους συμμαθητές του ήταν ευγενική* (His attitude towards his classmates was polite). Within the analysed dataset, AI-generated examples, particularly from GPT-5.3, were more balanced than in verbs, successfully covering multiple meanings in several cases (e.g. *κρίση* as both "crisis" and "judgment"). However, they remained less diverse and more predictable than Web-based data, tending to favor prototypical and pedagogically "safe" contexts rather than semantically rich or context-dependent uses. DeepSeek-V3 again showed acceptable performance in literal meanings but struggled with abstraction and domain variation.

As shown in Table 2, within the analysed dataset, GPT-5.3 achieved the highest overall scores in this category, followed by DeepSeek-V3 and Web-based examples.

Evaluation of Idiomatic Expressions

Idiomatic expressions proved to be the most challenging category, particularly for AI-generated data. Web-based examples generally showed high typicality, strong contextual grounding, and natural variation in meaning. For instance:

Table 2

Mean Evaluation Scores for Polysemous Nouns

Source	Typicality	Informativeness	Intelligibility	Pedagogical Appropriateness	Mean Total
Web	1.9	1.8	1.6	1.6	6.9
GPT-5.3	1.8	1.8	2.0	1.8	7.4
DeepSeek-V3	1.7	1.7	2.0	1.7	7.1

Note. Κρίση [krísi], στάση [stási], γλώσσα [glósa].

Table 3

Mean Evaluation Scores for Idiomatic Expressions

Source	Typicality	Informativeness	Intelligibility	Pedagogical Appropriateness	Mean Total
Web	1.9	1.8	1.7	1.8	7.2
GPT-5.3	1.6	1.7	1.9	1.6	6.8
DeepSeek-V3	0.9	1.2	1.6	1.0	4.7

Note. Τρώγομαι με τα ρούχα μου [trógome me ta rúha mu], βγάζει μάτι [vgázi máti], λέω ένα χεράκι [léo éna heráki].

Δεν τρώγομαι με τα ρούχα μου, είμαι πολύ καλά. (I don't let things get to me; I'm doing very well.) This example reflects the idiomatic meaning of emotional restraint or internal tension in a natural discourse context. GPT-5.3 showed moderate success, producing generally valid and intelligible examples, such as: *Όταν περιμένω να γράψουμε τεστ, τρώγομαι με τα ρούχα μου γιατί έχω άγχος* (When I'm waiting for a test, I get restless because I feel anxious). While acceptable, such examples tend to simplify the idiomatic meaning toward anxiety or nervousness, reducing semantic nuance. In contrast, DeepSeek-V3 exhibited systematic failures, including literalization and blending of distinct expressions: *Αν πούμε ένα χεράκι όλοι μαζί, θα τελειώσουμε πιο γρήγορα* (If we all "say a little hand" together, we will finish the group project in a few minutes.). This example was evaluated as confusion between the idioms «λέω ένα χεράκι» and «δίνω ένα χεράκι» ("give/lend a hand"). Similarly, literal reinterpretations were observed: *Αν συνεχίσεις να τρώγεις με τα ρούχα σου, θα τα χαλάσεις* (If you keep 'eating yourself with your clothes', you'll ruin things [incorrect literal interpretation]). Here, the idiom is incorrectly interpreted in a literal sense, leading to semantic inconsistency. Table 3 summarizes the performance differences across sources.

Within the analysed dataset, AI-generated examples received lower evaluation scores for idiomatic expressions than for the other lexical categories, particularly in the case of DeepSeek-V3.

Comparative Synthesis Across Categories

To synthesize the comparative findings across lexical categories, Table 4 presents an overview of the relative performance of Web-based and AI-generated examples in terms of pedagogical suitability.

As shown in Table 4, Web-based examples received higher scores for semantic coverage in polysemous verbs and idiomatic expressions, where figurative and context-dependent meanings were more frequently represented, at the cost of higher contextual complexity. AI-generated examples, particularly from GPT-5.3, showed greater structural uniformity and higher intelligibility, most notably for polysemous nouns. DeepSeek-generated examples received lower scores across categories, including non-standard collocations and literal reinterpretations of idiomatic meaning.

Students' Evaluation

A total of 41 lower secondary school students evaluated 27 selected usage examples in total, corresponding to nine target lexical items. For each lexical item, students evaluated one Web-based example, one GPT-5.3-generated example, and one DeepSeek-V3-generated example. As shown in Table 5, overall differences among the three sources were minor. Web-based examples received slightly higher ratings in correct meaning identification, as well as in meaning comprehension, clarity, usefulness, and preference compared to GPT-5.3 and DeepSeek-V3 examples, which received nearly identical ratings. It should be noted that preference percentages do not sum to 100%, as a small proportion of students did not indicate a preference for any of the three examples presented.

However, the results differed across lexical categories, indicating that students' performance was not uniform. As shown in Table 6, Web-based examples were consistently

preferred for idiomatic and figurative uses, receiving the highest preference ratings in this category (34.00%). For polysemous nouns, DeepSeek-generated examples received the highest preference rating (29.68%), while GPT-5.3 and Web examples performed comparably. For polysemous verbs, Web-based examples received the highest preference (33.43%), with DeepSeek-V3 second (32.40%) and GPT-5.3 third (31.34%). Several item-level findings are worth noting. For the noun *στάση*, GPT-5.3 received a markedly high preference rating of 63.4%, the highest recorded for any single item across all sources and categories. AI-generated examples may be perceived as particularly accessible by learners. A similarly strong preference for GPT-5.3 was observed for the idiomatic expression *βγάζει μάτι* (61%), an item for which GPT-5.3 produced valid and contextually appropriate examples. By contrast, item-level results for idiomatic expressions revealed a markedly different pattern. For the idiom *τρώγομαι με τα ρούχα μου*, correct meaning identification rates dropped substantially across all AI-generated sources, with DeepSeek-generated examples yielding the weakest performance in the entire idiomatic category. These item-level results indicate considerable variability among lexical items within the same broad lexical category.

In general, students rated all three sources similarly in terms of perceived clarity and usefulness, with Likert-scale ratings ranging narrowly between 3.90 and 4.00 across sources. However, correct meaning identification rates showed a clearer differentiation, with Web-based examples (56.50%) over GPT-5.3-generated (54.26%) and DeepSeek-generated examples (51.54%). The largest difference

Table 4

Synthesis of Comparative Findings Across Lexical Categories

Category	Best-performing Source	Strength	Key limitation
Polysemous verbs	Web	Extensive semantic and figurative range	Higher contextual complexity
Polysemous nouns	GPT-5.3	High clarity and pedagogical suitability	Limited variation compared to Web data
Idioms	1. Web 2. GPT-5.3 (partial)	High authenticity and pragmatic accuracy Intelligible baseline examples	Requires expert filtering for pedagogical use Reduced nuance; DeepSeek-V3 shows recurrent low-scoring outputs

Table 5

Mean Evaluation Scores from Students' Evaluation Across Sources

Source	Correct meaning (%)	Meaning	Clarity	Usefulness	Preference
Web	56.50%	4.00	3.93	3.95	32.00%
GPT-5.3	54.26%	3.98	3.91	3.93	31.50%
DeepSeek-V3	51.54%	3.95	3.90	3.91	31.81%

Note. Correct meaning identification was used as an objective measure of comprehensibility, while Likert-scale ratings (meaning comprehension, clarity, and usefulness) captured perceived effectiveness. Preference responses indicated which examples students found most helpful overall.

Table 6*Mean Evaluation Scores from Students' Evaluation by Lexical Category and Source*

Category	Source	Mean % correct	Meaning	Clarity	Usefulness	Preference
Verbs	Web	72.14%	4.20	4.02	4.11	33.43%
	GPT-5.3	72.88%	4.24	4.13	4.15	31.34%
	DeepSeek-V3	68.61%	4.27	4.18	4.18	32.40%
Nouns	Web	60.33%	4.03	3.95	3.91	26.45%
	GPT-5.3	56.85%	4.03	3.97	3.91	25.63%
	DeepSeek-V3	48.83%	4.01	3.95	3.91	29.68%
Idiomatic expressions	Web	40.77%	3.67	3.66	3.73	34.00%
	GPT-5.3	32.37%	3.59	3.63	3.65	29.27%
	DeepSeek-V3	31.60%	3.64	3.65	3.63	33.67%

Note. Preference percentages within each category do not sum to 100% as some students did not express a preference.

between perceived and actual comprehension performance was observed in the idiomatic expressions category, where Web-based examples achieved a correct identification rate of 40.77% compared to 31.60% for DeepSeek-V3, despite relatively similar clarity ratings across sources. While DeepSeek-V3-generated examples received marginally lower meaning comprehension and clarity ratings than the other two sources across all categories, this difference was most pronounced in the idiomatic expressions category, where the gap between sources in correct meaning identification was also the largest.

DISCUSSION

The study's findings are situated within the broader literature on Web-based and AI-assisted lexicography, addressing each research question in turn. Regarding the first research question, how Web-based and AI-generated usage examples compare in terms of pedagogical suitability across different categories of lexical complexity, the findings suggest category-dependent tendencies within the analysed dataset. The results indicate that the effectiveness of each source is not uniform but closely linked to the type of lexical complexity involved, a pattern that emerged consistently across both the expert evaluation and the learner validation phase. This category-sensitive perspective extends existing models of selecting examples; however, it is proposed that pedagogical appropriateness should not be evaluated only through general criteria (i.e., authenticity, clarity and typicality), but also in the context of the lexical and semantic demands of the target item. Within the analysed dataset, Web-based examples generally indicated greater semantic diversity and contextual richness, particularly in cases involving polysemous verbs and idiomatic expressions. These examples reflected

contemporary language use and captured a wide range of figurative and context-dependent meanings, aligning with previous research on the value of Web-derived linguistic data for lexicographic purposes (Kilgarriff, 2007; Fuertes-Olivera, 2012; Davies, 2018; Gatto, 2025). At the same time, however, their pedagogical usability was not always straightforward. Increased contextual complexity, domain-specific references, and stylistic variation often required adaptation to ensure accessibility for younger learners. This tension between authenticity and accessibility reflects a well-documented challenge in pedagogical lexicography, where the richness of naturally occurring data must be balanced against the cognitive and linguistic needs of the target learner group (Atkins & Rundell, 2008; Tarp & Gouws, 2020).

AI-generated examples showed a different distribution of strengths. GPT-5.3 generally produced grammatically correct and intelligible examples and performed particularly well in cases where meanings are more stable and can be expressed in clear, prototypical contexts, such as polysemous nouns. In these cases, AI outputs achieved a balance between clarity and pedagogical appropriateness. However, their semantic coverage was often limited, especially for lexical items characterized by extensive polysemy or idiomatic variation. These findings are consistent with previous studies indicating that large language models perform well in generating structurally coherent sentences but may struggle with idiomatic precision and nuanced semantic variation (Phoodai & Rikk, 2023; Jakubiček & Rundell, 2023; Cai et al., 2024; Almeman et al., 2024; Merx et al., 2024; Arnett & Bergen, 2025). It should be acknowledged that these observations reflect performance under a specific zero-shot prompting configuration; different results might be obtained under few-shot or expert-guided prompting conditions, and the present findings cannot be taken as definitive assessments of LLM capability.

The limitations of AI-generated examples became more pronounced in the domain of idiomatic expressions. While GPT-5.3 produced acceptable baseline examples, these often showed reduced contextual diversity and occasional semantic simplification. DeepSeek-generated examples demonstrated more substantial limitations, including frequent misinterpretations, collocational errors, and shifts toward literal meaning. These findings suggest that model performance varies, with idiomatic language remaining a particularly challenging domain for LLMs, especially in morphologically rich and less-resourced languages (Abdelrahman, 2024; Beliga & Filipović Petrović, 2024; Arnett & Bergen, 2025; Marković & Stanković, 2025; Zhong et al., 2026;). The weaker performance of DeepSeek across all categories may partly reflect differences in training data coverage for Modern Greek, though this cannot be confirmed without access to model documentation. It may also reflect differences in prompting sensitivity between models. These alternative explanations should be considered alongside any interpretation of the results. The findings also suggest that source effectiveness may be both item-specific and category-dependent. For example, GPT-5.3 performed much better at providing contextual usage for the idiomatic expression “βγάζει μάτι” than for “τρώγομαι με τα ρούχα μου”. This suggests that broad lexical categories do not fully predict pedagogical suitability. Other factors, such as semantic transparency, conventionality of meaning, contextual familiarity, and learners’ interpretation may also influence whether a usage example works for pedagogical purposes.

A more informative picture emerges when correct meaning identification rates are considered alongside perceived ratings. Across all categories, Web-based examples consistently yielded the highest correct identification rates, despite not always receiving the highest clarity or usefulness ratings from students. This gap between perceived and actual comprehension is particularly pronounced in the idiomatic expressions category, where Web-based examples achieved a correct identification rate of 40.77% compared to 31.60% for DeepSeek-V3, even though subjective ratings across sources remained relatively close. This pattern is consistent with research demonstrating that authentic, contextually grounded examples support deeper semantic processing, even when learners do not explicitly recognize their superiority (Tarp, 2014; Kosem et al., 2019) and underscores that learner preference alone is an insufficient indicator of pedagogical quality.

The findings provide support for all three hypotheses formulated at the outset of the study, albeit with important qualifications. The first hypothesis, that Web-based examples would demonstrate greater semantic variability and contextual richness, particularly for figurative and idiomatic language, was supported within the analysed dataset, with Web-derived examples consistently achieving higher typicality and informativeness scores, and higher correct meaning identification rates in the learner evaluation. The second hypothesis, that AI-generated examples would show higher

intelligibility and structural consistency, was also supported, with GPT-5.3 achieving the highest intelligibility scores across categories, a pattern that translated into stronger learner preference for semantically stable items such as polysemous nouns. The third hypothesis, that source performance would vary systematically across lexical categories rather than being uniform, received the strongest support from the data. The advantage of Web-based examples was most pronounced for idiomatic expressions, where AI-generated outputs showed the most substantial limitations, while GPT-5.3 performed comparably to or above Web-based examples for polysemous nouns. This category-dependent pattern is the most theoretically significant finding of the study, as it suggests that the choice of data source in pedagogical lexicography cannot be made independently of the type of lexical complexity involved.

With respect to the second research question, what patterns of variation emerge across lexical categories and what implications these have for the integration of different data sources in lexicographic practice, the findings point to several methodological and pedagogical considerations that bear directly on how hybrid workflows might be designed. Web-based examples require extensive filtering and adaptation to ensure clarity, appropriateness, and alignment with learner needs. The retrieval process is also time-consuming, particularly in languages like Modern Greek, where morphological variation increases search complexity (Hoenen et al., 2020). Furthermore, the use of a general-purpose search engine as a retrieval tool introduces variability in results across time and search configurations, raising questions about reproducibility that would need to be addressed in larger-scale lexicographic workflows. Therefore, the quantity of web-based usage examples that have been observed should be interpreted as an outcome of specific search, filtering and selection procedures, and not a stable aspect of the entire category of web material. Additional challenges arise from the nature of Web-based content itself. Authentic examples sourced from journalistic and informational websites may contain politically sensitive references, domain-specific terminology, or culturally embedded content that requires careful review before inclusion in materials designed for school-aged learners. Several examples retrieved in the present study (particularly those involving geopolitical events) illustrate this tension between currency and age-appropriateness, underscoring the need for systematic pedagogical filtering beyond linguistic criteria alone.

AI-generated examples present a distinct set of integration challenges, as also noted in recent research on AI-assisted Modern Greek (MG) lexicography (Alexandri & Iordanidou, 2025). While they offer efficiency and structural consistency, they require systematic validation to ensure semantic accuracy, idiomatic naturalness, and pedagogical suitability (Nasution & Onan, 2024). The lack of transparent sourcing further raises questions about authenticity in lexicographic contexts,

as it is not possible to verify whether generated examples reflect actual patterns of language use or are artefacts of the model's training data distribution. This concern is particularly relevant for idiomatic expressions, where subtle collocational conventions and pragmatic norms may not be reliably reproduced by current models, especially in languages that are less well-represented in LLM training corpora.

Taken together, these findings highlight a complementary relationship between the two data sources. Web-based examples provide authenticity, semantic breadth, and access to evolving language use, while AI-generated examples offer structural clarity, consistency, and efficiency. However, neither source can independently meet the pedagogical and lexicographic requirements of a learner-oriented dictionary. These results therefore support the development of hybrid lexicographic approaches that combine Web-derived data, AI-assisted generation, and expert curation (de Schryver, 2024; Rundell, 2024). In such a model, Web-based examples can serve as a source of authentic linguistic input, while AI-generated examples can be used to produce simplified, pedagogically adapted variants for semantically stable items. Crucially, human expertise remains central in ensuring accuracy, selecting appropriate examples, and aligning lexicographic content with learners' cognitive and educational needs. These findings align with previous research emphasizing the importance of human oversight in AI-assisted lexicography (Lew, 2023; Rundell, 2024; Kosem et al., 2024; Lew, 2024), while also extending this discussion by demonstrating that the relative utility of different data sources varies systematically across lexical categories, a distinction that has implications for how hybrid workflows should be designed and prioritized in practice. More broadly, the findings of this study suggest that pedagogical example selection may need to be reconceptualized as a category-sensitive process. What constitutes a "good" usage example for a Modern Greek pedagogical dictionary cannot be assessed only in terms of general qualities such as authenticity, clarity, or typicality, but also in relation to lexical complexity, pedagogical function, and learners' ability to interpret meaning in context. Although this reconceptualization may be viewed as tentative due to the exploratory design of this research study, it provides a constructive direction for future research and for the development of more differentiated frameworks in pedagogical lexicography.

Limitations

Several methodological limitations should be considered when interpreting the findings of this study. First, the study is exploratory and is based on a purposive set of 81 usage examples. The selected lexical items functioned as case studies of different types of lexical complexity. Second, the learner validation phase involved 41 middle school students, which limits the extent to which the findings can be generalized to other learner groups. The analysis was descriptive and did

not include inferential statistical testing, meaning that the variations across sources and lexical categories should be interpreted as suggestive tendencies within this dataset rather than statistically generalisable effects. Third, the Web-based examples were collected through Google searches, which may be influenced by ranking algorithms, changes over time, and personalization processes. The decision to restrict the dataset to the first three relevant results per query may have also introduced a selection bias, privileging more dominant discourse types. Finally, the AI-generated examples were produced using one prompting configuration and one generation for each lexical item and model. While this design was chosen to approximate typical user interaction conditions, it does not show the full potential or variability of LLM-generated output under different prompting strategies or repeated generations.

CONCLUSION

This exploratory study examined the pedagogical suitability of Web-based and AI-generated usage examples for a Modern Greek learner dictionary targeting students aged 11-15. Within the analysed dataset, the findings suggested that pedagogical suitability was influenced not only by source type, but also by the type of lexical complexity involved. Web-based examples generally showed greater semantic diversity and stronger performance in figurative and idiomatic uses, while AI-generated examples, particularly those produced by GPT-5.3, tended to offer advantages in clarity and structural consistency, especially for semantically stable lexical items such as polysemous nouns.

The learner evaluation further suggested that perceived clarity does not always correspond to successful identification of meaning. While students often rated the three sources closely in terms of clarity and usefulness, Web-based examples achieved higher rates of correct identification of meaning (particularly in figurative and idiomatic uses). These findings highlight the importance of evaluating pedagogical suitability not only through perceived accessibility, but also through actual learner comprehension.

The current findings add support to a category-sensitive view, where the suitability of a usage-example source depends partly on the lexical and pedagogical demands of the task. Web-based and AI-generated examples should therefore not be treated as uniformly suitable or unsuitable. The study suggests that useful or appropriate usage appears to depend on the lexical and semantic properties involved in the target item. The findings also highlight the importance of expert mediation in assessing and choosing a "pedagogically usable example".

The study remains exploratory in scope and is based on a limited and purposive dataset. The analysed lexical items func-

tioned as case studies rather than a representative sample of Modern Greek vocabulary, and the learner validation reflected the responses of a limited and relatively homogeneous user group. The findings should therefore be seen as indicative tendencies in the sample analysed rather than as statistically generalisable effects. Future research should expand the lexical dataset, examine additional categories of lexical complexity (including collocations, register-specific vocabulary, and neologisms) and investigate how different prompting strategies, retrieval methods, and hybrid workflows may affect pedagogical suitability in lexicographic practice. While the analysis focused on Modern Greek, similar questions may also arise in other morphologically rich and relatively under-resourced languages. More broadly, the study points to the importance of category-sensitive and human-mediated approaches to evaluating pedagogical suitability and integration of usage examples in pedagogical lexicography.

USE OF AI TOOLS STATEMENT

The authors confirm that AI tools were used for language editing of this manuscript. All interpretations, analyses, and final wording remain the sole responsibility of the authors.

REFERENCES

- Abdelrahman, M. (2024). 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(12), 17-24. <https://polarpublications.com/index.php/JABADP/article/view/2024-12-10>
- Alexandri, K. (2025). Τεχνητή Νοημοσύνη και Λεξικογραφία: Μπορεί ένα διαλογικό ρομπότ να αντικαταστήσει το παιδαγωγικό λεξικό; [Artificial intelligence and lexicography: Can a chatbot replace the pedagogical dictionary?]. In I. Spantidakis, K. Ntinias, V. Chatzinikita, & E. Griva (Eds.), *Language, education and artificial intelligence* (pp. 159-174). University of Crete.
- Alexandri, K., & Iordanidou, A. (2025). Τεχνητή Νοημοσύνη και λεξικογραφία: Η συμβολή της ΤΝ στην επίλυση γλωσσικών αποριών [Artificial intelligence and lexicography: The contribution of AI to solving language-related queries]. *Studies in Greek Linguistics*, 44, 33-44. https://www.eshop.ins-auth.gr/images/companies/1/archive/MEG_PLIRI/MEG_44_33_44.pdf
- Almeman, F. Y., Schockaert, S., & Espinosa Anke, L. (2024). WordNet under scrutiny: Dictionary examples in the era of large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 17683-17695). European Language Resources Association and ICCL. <http://dx.doi.org/10.63317/4skerayorcxh>.
- Arkhangelskiy, T. (2019). Corpus of usage examples: What is it good for? In A. Arppe, J. Good, M. Hulden, J. Lachler, A. Palmer, L. Schwartz, & M. Silfverberg (Eds.), *Proceedings of the 3rd Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)* (pp. 56-63). Association for Computational Linguistics. <http://dx.doi.org/10.33011/computel.v1i.411>
- Arnett, C., & Bergen, B. (2025). Why do language models perform worse for morphologically complex languages? In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 6607-6623). Association for Computational Linguistics. <https://aclanthology.org/2025.coling-main.441/>
- Atkins, B. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press. <http://dx.doi.org/10.1093/oso/9780199277704.001.0001>
- Beliga, S., & Filipović Petrović, I. (2024). Large language models supporting lexicography: Conceptual organization of Croatian idioms. In Š. Arhar Holdt & T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities* (pp. 23-46). Ljubljana University Press. <http://dx.doi.org/10.5281/zenodo.13912515>
- Cai, B., Clarence, N., Liang, D., & Hotama, S. (2024). Low-cost generation and evaluation of dictionary example sentences. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Lan-*

DATA AVAILABILITY STATEMENT

All sets of data analysed throughout the current study, along with the results of learner validation questionnaires, are available from the authors upon request.

DECLARATION OF COMPETING INTEREST

None declared.

AUTHORS' CONTRIBUTIONS

Lamprini Lourida: data curation; formal analysis; investigation; methodology; resources; visualization; writing - original draft.

Katerina Alexandri: conceptualization; investigation; methodology; resources; supervision; writing - review and editing.

- guage Technologies, Volume 1: Long Papers* (pp. 3538-3549). Association for Computational Linguistics. <https://aclanthology.org/2024.naacl-long.194.pdf>
- Chi, A. (2022). Researching pedagogical lexicography. In H. Jackson (Ed.), *The Bloomsbury Handbook of Lexicography* (2nd ed., pp. 145-164). Bloomsbury Publishing. <https://hdl.handle.net/1783.1/116725>
- Crosthwaite, P., & Baisa, V. (2023). Generative AI and the end of corpus-assisted data-driven learning? Not so fast! *Applied Corpus Linguistics*, 3(3), Article 100066. <http://dx.doi.org/10.1016/j.acorp.2023.100066>
- Davies, M. (2018). Corpus-based studies of lexical and semantic variation: The importance of both corpus size and corpus design. In A. Meurman-Solin, C. Seoane, & M. José López-Couso (Eds.), *From data to evidence in English language research* (pp. 66-87). Brill. http://dx.doi.org/10.1163/9789004390652_004
- de Schryver, G.-M. (2024). The road towards fine-tuned LLMs for lexicography. In S. Krek (Ed.), *Book of abstracts of the Workshop "Large Language Models and Lexicography"*, 8 October 2024, Cavtat, Croatia (pp. 6-11). ELEXIS Association. <http://hdl.handle.net/1854/LU-01JVR7H9DE5TTXV4GPNEW30PC>
- Durward, M., & Thomson, C. (2024). Evaluating vocabulary usage in LLMs. In E. Kochmar, M. Bexte, J. Burstein, A. Horbach, R. Laarmann-Quante, A. Tack, V. Yaneva, & Z. Yuan (Eds.), *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)* (pp. 266-282). Association for Computational Linguistics. <https://aclanthology.org/2024.bea-1.22/>
- Fuertes-Olivera, P. A. (2012). Lexicography and the Internet as a (re-)source. *Lexicographica*, 28, 49-70. <http://dx.doi.org/10.1515/lexi.2012-0005>
- Garipov, T., Morozov, D., Gubarkova, Y., Kozerenko, A., & Glazkova, A. (2026). An experimental study of automating explanatory dictionary compilation with language models. In M. Bakaev et al. (Eds.), *Internet and modern society: IMS 2025 (Communications in Computer and Information Science, Vol. 2671)*, pp. 117-131. Springer. http://dx.doi.org/10.1007/978-3-032-04958-2_9
- Gatto, M. (2025). Web as corpus for data-driven learning. In *The Palgrave Encyclopedia of Computer-Assisted Language Learning* (pp. 1-8). Springer Nature Switzerland. http://dx.doi.org/10.1007/978-3-031-51447-0_66-1
- Gavriilidou, M., Lampropoulou, P., & Angelakos, K. (2007). *Ερμηνευτικό λεξικό νέας ελληνικής Α', Β', Γ' Γυμνασίου* [Explanatory dictionary of modern Greek for grades 7, 8, and 9 of secondary school]. OEDB.
- Hoenen, A., Koc, C., & Rahn, M. D. (2020). A manual for Web corpus crawling of low resource languages. *Umanistica Digitale*, 4(8). <http://dx.doi.org/10.6092/issn.2532-8816/9931>
- Ide, Y., Nohejl, A., Tanner, J., Yanaka, H., Lindsay, C., & Watanabe, T. (2026). Towards automated lexicography: Generating and evaluating definitions for learner's dictionaries. arXiv. <http://dx.doi.org/10.48550/arXiv.2601.01842>
- Iordanidou, A., Mantzari, E., & Pantazara, M. (Eds.). (2007). *Λεξικό της ελληνικής ως ξένης γλώσσας για μαθητές της δευτεροβάθμιας εκπαίδευσης* [Dictionary of Greek as a foreign language for secondary school students]. National and Kapodistrian University of Athens.
- Jakubiček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for postediting lexicography? In M. Medved, M. Měchura, I. Kosem, J. Kallas, C. Tiberius, & M. Jakubiček (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference* (pp. 518-533). Lexical Computing CZ s.r.o. <https://elex.link/ojs/index.php/elex/article/view/46/30>
- Kapsalis, G., Paskhalis, A., Tsiolos, S., & Tsioulis, K. (2007). *Ορθογραφικό-ερμηνευτικό λεξικό Δ', Ε', ΣΤ' Δημοτικού* [Orthographic and explanatory dictionary for grades 4, 5, and 6 of primary school: Our dictionary]. OEDB.
- Kilgarriff, A. (2007). Googleology is bad science. *Computational Linguistics*, 33(1), 147-151. <http://dx.doi.org/10.1162/coli.2007.33.1.147>
- Kilgarriff, A., & Grefenstette, G. (2003). Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3), 333-347. <http://dx.doi.org/10.1162/089120103322711569>
- Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In A. DeCesaris & E. Bernal (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 425-432). Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra. https://euralex.org/elx_proceedings/Euralex2008/
- Kosem, I., Gantar, P., Holdt, Š. A., Gapsa, M., Zgaga, K., & Krek, S. (2024). AI in lexicography at the University of Ljubljana. In S. Krek (Ed.), *Book of abstracts of the workshop "Large Language Models and Lexicography"*, 8 October 2024, Cavtat, Croatia (pp. 29-32). ELEXIS Association. https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf#page=11
- Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J., & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: The case(s) of GDEX. *International Journal of Lexicography*, 32(2), 119-137. <http://dx.doi.org/10.1093/ijl/icy014>
- Krstev, C., & Stanković, R. (2023). Language report Serbian. In G. Rehm & A. Way (Eds.), *European language equality*. (pp. 203-206). Springer. http://dx.doi.org/10.1007/978-3-031-28819-7_32

- Laippala, V., Egbert, J., Biber, D., & Kyröläinen, A.-J. (2021). Exploring the role of lexis and grammar for the stable identification of register in an unrestricted corpus of web documents. *Language Resources and Evaluation*, 55(3), 757-788. <http://dx.doi.org/10.1007/s10579-020-09519-z>
- Lew, R. (2015). Research into the use of online dictionaries. *International Journal of Lexicography*, 28(2), 232-253. <http://dx.doi.org/10.1093/ijl/ecv010>
- Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10, Article 704. <http://dx.doi.org/10.1057/s41599-023-02119-6>
- Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11, Article 426. <http://dx.doi.org/10.1057/s41599-024-02889-7>
- Lewandowska-Tomaszczyk, B., & Pawłowski, G. (2025). Testing ChatGPT on terminology generation, definitions: Translation, and ontology creation in German, English and Polish. *Research in Language*, 23, 320-340. <http://dx.doi.org/10.18778/1731-7533.23.20>
- Marković, A. & Stanković, R. (2025). So close but still far: Case study on application of LLMs in idioms identification, definition, and generation of illustrative examples. In I. Kosem, M. Jakubiček, M. Medved', K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic lexicography in the 21st Century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 Conference* (pp. 79-94). Lexical Computing.
- Merx, R., Vylomova, E., & Kurniawan, K. (2024). Generating bilingual example sentences with large language models as lexicography assistants. In T. Baldwin, S. J. Rodríguez Méndez, & N. Kuo (Eds.), *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association* (pp. 64-74). Association for Computational Linguistics. <https://aclanthology.org/2024.alta-1.5/>
- Nasution, A. H., & Onan, A. (2024). ChatGPT label: Comparing the quality of human-generated and LLM-generated annotations in low-resource language NLP tasks. *IEEE Access*, 12, 71876-71900. <http://dx.doi.org/10.1109/access.2024.3402809>
- Phoodai, C., & Rikk, R. (2023). Exploring the capabilities of ChatGPT for lexicographical purposes: A comparison with Oxford Advanced Learner's Dictionary within the microstructural framework. In I. Kosem, M. C. Culy, M. Gantar, J. Kallas, S. Krek, & H. Tuul (Eds.), *Electronic lexicography in the 21st Century (eLex 2023): Proceedings of the eLex 2023 Conference* (pp. 345-375). Lexical Computing CZ. <https://elex.link/ojs/index.php/elex/article/view/35>
- Rundell, M. (2024). Automating the creation of dictionaries: Are we nearly there? *Humanising Language Teaching*, 26(1).
- Rundell, M., Jakubiček, M., Kovář, V., Matuška, O., & Cukr, M. (2025). Lexicom at 25: Reflections on the changing world of lexicography and language technology. In I. Kosem, M. Jakubiček, M. Medved', K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 Conference* (pp. 136-149). Lexical Computing CZ. https://elex.link/elex2025/wp-content/uploads/eLex2025-09-Rundell_etal.pdf
- Tarp, S. (2014). Dictionaries in the internet era: Innovation or business as usual? Enrique Alcaraz Memorial Lecture 2014. *Revista Alicantina de Estudios Ingleses*, 27, 233. <http://dx.doi.org/10.14198/raei.2014.27.13>
- Tarp, S., & Gouws, R. (2020). Reference skills or human-centered design: Towards a new lexicographical culture. *Lexikos*, 30, 470-498. <http://dx.doi.org/10.5788/30-1-1600>
- Vakalopoulou, A. (2000). *Το πρώτο μου λεξικό για το Δημοτικό* [My first dictionary for primary school]. Patakis Editions.
- Vakalopoulou, A., & Iordanidou, A. (2001). *Το λεξικό του Δημοτικού* [The dictionary of primary school]. Patakis Editions.
- Verlinde, S., & Binon, J. (2009). Pedagogical lexicography revisited. In H. Bergenholtz, S. Nielsen, & S. Tarp (Eds.), *Lexicography at a crossroads: Dictionaries and encyclopedias today, lexicographical tools tomorrow* (pp. 69-89). Peter Lang.
- Zhong, A. (2025). Prompts for language learners: A practical guide to using DeepSeek as a dictionary. *Lexikos*, 35(1), 274-285. <http://dx.doi.org/10.5788/35-1-2037>.
- Zhong, T., Yang, Z., Liu, Z., Zhang, R., You, W., Liu, Y., Sun, H., Pan, Y., Li, Y., Zhou, Y., Jiang, H., Chen, J., Li, X., & Liu, T. (2026). Opportunities and challenges of large language models for low-resource languages in humanities research. *arXiv*. <http://dx.doi.org/10.48550/arXiv.2412.04497>