https://doi.org/10.5507/magister.2025.514

Educational Implications of Lexicographic Misalignment: AI-Generated Definitions of English Loanwords in L2 Context

Jakov Proroković

Abstract: By replicating the experimental design of Balenović and Proroković (2025), on the same lexical dataset, the present research investigates whether the GPT-5 model yields more accurate, contextually appropriate, and linguistically differentiated lexicographic output when it comes to senses and usages of English loanwords in Croatian. Additionally, the aim is to determine how GPT-5's definitions, contextual examples, and sense partitioning compare qualitatively and quantitatively to those produced by GPT-4o. Despite improvements, the findings indicate that the GPT-5 model continues to display systematic vulnerabilities in L2 contexts, particularly for low-frequency loanwords. Though not exclusively, as in GPT-4o, less attested items elicit a higher rate of L1-to-L2 overgeneralization, leading to unverified or semantically implausible uses presented with high confidence. The study concludes that, although GPT-5 represents a significant step forward in the lexicographic endeavor, its performance remains inconsistent for L2 loanword interpretation. In other words, there still seems to exist the continued need for critical human oversight and the educational importance of AI literacy. Model's misinterpretation of prompt design and intent entails that recognizing and correcting such errors requires not only user awareness but language competence as well.

Keywords: *large language models (LLMs), GPT-5, cross-linguistic lexicography, loanword sense definitions, overgeneralization and AI overconfidence*

1 Introduction

Lexicography has long been regarded as a clearly delineated and methodologically stable branch of linguistics. At its core, the lexicographic endeavor involves synthesizing attested linguistic evidence across diverse

registers and contexts of use, identifying recurrent semantic and formal regularities, and systematizing these into structured entries that represent a term's meanings and realizations in relation to its defining traits (Bergenholtz & Tarp, 2003; Atkins & Rundell, 2008). In other words, lexicography mediates between raw linguistic data and organized lexical knowledge. However, this apparent methodological straightforwardness becomes more complex when applied within the domain of contact linguistics, where linguistic systems interact, borrow, and influence one another dynamically. In such contexts, the stability of lexical meaning, often presupposed in traditional lexicographic frameworks, is challenged by processes of borrowing, semantic shift, and hybridization. The present study therefore investigates how lexicographic principles can be adapted or reconceptualized for contact linguistic settings, and how artificial intelligence (AI)—particularly large language models (LLMs)—can assist or transform such analysis. More precisely, it explores whether the most recent AI models can avoid the unnecessary (either unattested or plain wrong) overgeneralization that was observed in earlier models when defining loanwords in their non-native contexts (see Balenović & Proroković, 2025).

1.1 Lexicography and AI

Research on large language models (LLMs) for linguistic and lexicographic purposes has expanded rapidly in recent years (cf. de Schryver, 2023; Lew, 2024; Chen et al., 2024; Klosa-Kückelhaus & Tiberius, 2025 etc.). Even the early models started to integrate contextualized representations that capture more finer-grained semantic nuances (cf. Ethayarajh, 2019; Yarbro & Olney, 2021, Liu et al., 2021), finally resulting in systems that demonstrated considerable potential for assisting in the generation of candidate definitions and sense inventories, thereby accelerating the discovery and initial drafting phases of lexicographic work (Balenović & Proroković, 2025; Lew, 2024). AI systems no longer merely replicate traditional dictionary-defining practices but increasingly engage in context-sensitive definitional reasoning, allowing for new forms of semi-automated lexical analysis, thus inviting a reconsideration of the epistemological and methodological foundations of lexicography in the age of intelligent computational models. Nevertheless, despite notable improvements in language-specific competencies and context sensitivity, the greatest danger lurks in unreservedly taking the AI's output without having in mind that the models tend to generate probability-based sequences of linguistic structures they extrapolated from

the training data, but without necessarily being referenced or grounded in meaning (where meaning implies extra-linguistic awareness), thus becoming a form of "stochastic parrots" (Bender et al., 2021).

What models tend to excel at is using distributional patterns and memorizing encyclopedic facts to produce definitions (fast, fluent), and with the more recent advancements capturing richer, distributional and contextual information than single-token embeddings alone. At the moment, it seems rather inevitable that the use of LLMs will reshape the field of lexicography, especially in the cross-lingusitic and cross-cultural context and in relation to learners' interaction with the emerging technology. In fact, this is already happening in the sense that many learners seem to be gradually shifting from relying on traditional print or electronic dictionaries toward querying LLMs (especially ChatGPT) as a first stop for meaning, usage, and collocations (see Ptasznik & Lew, 2025; Liu et al., 2025). In other words, it is now perfectly clear that we can (if not ought to) use them for unearthing candidate definitions and sense inventories, for they significantly accelerate discovery and initial drafting. This does not entirely eliminate dictionary use, but rather reorders it: learners often turn to dictionaries only when the LLM fails, or to verify or expand on LLM outputs. In fact, learners sometimes detect errors or hallucinations in LLM output and revert to dictionaries as a sanity check, so the relationship is not simple replacement but dynamic interplay (Liu et al., 2025). In other words, LLMs are reshaping dictionary use patterns among language learners by inserting themselves as an intermediate interpretive tool, with learners treating LLMs as "first-instance interpreters," and dictionaries becoming secondary resources (ibid.).

Though lexicographic in nature, this study also carries broader implications for the development of users' critical awareness when interacting with AI systems. In particular, it highlights the necessity for users to reflect consciously on what they seek to find out and how they formulate their queries; an issue repeatedly identified in recent educational research as a persistent challenge for learners and educators alike (Woo et al., 2024; Milakis et al., 2025; Kang et al., 2025; Kim et al., 2025). A substantial part of employing AI in dictionary compilation lies in the careful design of prompts. As the results of this study (or at least its implications) will illustrate, effective prompting is neither trivial nor mechanical for it requires a high degree of linguistic awareness, disciplinary literacy, and attention to contextual and semantic detail. Moreover, as de Schryver (2023) cautions, while replacing a series of semi-automated lexicographic tasks (such as corpus building,

annotation, or headword list generation) with a single well-crafted prompt represents a remarkable technical advance, it does not remove the human element required for the framing of the prompt that directly conditions the quality and interpretability of the AI output. Large language models produce non-deterministic responses and even identical prompts may yield divergent results, thus underscoring the need for human oversight and reflective engagement to ensure that AI-generated material aligns with lexicographic standards and communicative intent (see de Schryver, 2023; Beliga & Filipović Petrović, 2024; Balenović & Proroković, 2025; Klosa-Kückelhaus & Tiberius, 2025 etc.).

1.2 On the Shoulders of the Previous Study

This study builds directly on the pioneering work of Balenović and Proroković (2025), *"The lexicographic potential of artificial intelligence: a case study of English loanwords in the Croatian language,"* in which the authors explored how generative AI models can distinguish senses and usages of English loanwords in Croatian. Their study showed that, while ChatGPT is quite capable of generating dictionary-style definitions for English loanwords used in Croatian (especially those that are frequent and well documented), its reliability diminishes for less frequent terms. In such cases the model tends to overgeneralize from English, importing senses that are not actually used in Croatian, and sometimes producing incorrect or unattested examples. Comparisons between early-2024 and early-2025 data revealed modest improvements, with the newer version of ChatGPT demonstrating greater nuance in dealing with ambiguous usages and somewhat better alignment with expected senses. Nevertheless, inconsistencies remained, particularly in how word frequency correlated with sense number, indicating that high frequency did not always yield more distinct senses, contrary to what one might expect.

Said findings set a foundation for the present study, which preserves the same sample of loanwords but uses an upgraded model (GPT-5), in order to test whether those improvements continue, and whether the limitations noted can be addressed. By adopting the same set of English loanwords as in Balenović & Proroković's (2025) for a representative sample, this paper ensures continuity with the same empirical footing, but pushes the inquiry further by testing a more advanced underlying architecture: the GPT-5 model. In this way, the work can be seen as a natural extension of the previous paper: it not only replicates their experimental design (on the

same lexical material), but also investigates how advancements in model capacity and contextual reasoning might improve, refine, or challenge their original findings. This continuity in sample selection enables valid comparison across model developmental periods, while the shift to GPT-5 allows us to test whether more powerful models can overcome the limitations documented in the earlier work. Finally, the goal of this paper can be summarized as an attempt to answer the question of the ways in which GPT-5's definitions, contextual examples, and sense delineations differ qualitatively and quantitatively from those generated by GPT-4o for the same set of items?

2 Study Design

2.1 Research Questions

Building on Balenović and Proroković's (2025) findings, and employing the same corpus of English loanwords attested in Croatian, this study seeks to examine how the newer GPT-5 model performs in comparison to GPT-4o. The focus is on assessing whether the model's enhanced contextual reasoning and linguistic depth translate into more accurate and context-sensitive lexicographic performance. Accordingly, the research is guided by the following questions:

1. How effectively does the model capture language-specific contextual differences in meaning and register behavior of English loanwords used in Croatian?
2. To what extent does GPT-5, compared to GPT-4o, distinguish between the semantic and pragmatic nuances of the same lexical item when used in English (L1) versus Croatian (L2)?
3. Are higher-frequency loanwords handled more consistently and precisely than low-frequency or emerging ones, and if so, how does this pattern compare to that observed in GPT-4o?

2.2 Method

The methodological framework of this study closely follows that established by Balenović and Proroković (2025), ensuring consistency and comparability between the two research stages. As in the previous study, prompts were used to elicit the model's definitions, categorizations, and contextual interpretations of English loanwords attested in Croatian. However, for the present investigation, only four prompts were employed, as opposed to the six used previously. Specifically, the model was queried using the following:

Senses of "X" in Croatian, *Word class of "X" in Croatian*, *Senses of "X" in English*, and *Word class of "X" in English*.

Table 1. Prompts designed to elicit information on the target English loanword where »X« designates the targeted loanword

Feedback goal	Prompt designed to address it
Senses of »X« in Cro	Ako postoje, koja su moguća značenja/načini uporabe riječi »X« u hrvatskom jeziku?
Word class of »X« in Cro	Kao koja vrsta riječi se riječ »X« može koristiti u hrvatskom jeziku?
Senses of »X« in Eng	What are the possible senses of the word »X« in English?
Word class of »X« in Eng	Which word classes can the word »X« represent when used in English?

This slight methodological reduction was introduced in order to focus on the model's ability to distinguish cross-linguistic semantic and grammatical differences between the same lexical item used in Croatian (L2) and English (L1), while maintaining a reproducible dataset. Each prompt was issued individually (with the new session being started each time) in identical conditions to ensure consistency, and responses were recorded and analyzed for accuracy, contextual adequacy, and internal consistency. This approach enables a direct comparison of GPT-5's output with that obtained from GPT-4o, providing insight into the model's development in lexicographic competence.

2.3 Sample and Data Collection

The lexical sample used in this study corresponds to that analyzed by Balenović and Proroković (2025), ensuring full comparability between the two datasets. For the exact list of English loanwords tested on the GPT-5 model, see Balenović and Proroković (2025). As in the previous research, the selection of items is grounded in the work of Bogunović and Kučić (2022) and Bogunović (2023), who compiled the ENGRI corpus—a corpus of English-origin words attested in Croatian. The sample focuses on loanwords that have largely retained their original orthographic and phonetic forms and that are verifiably integrated into contemporary Croatian usage, thus ensuring relevance and representativeness across frequency levels. For the purposes of the present analysis, all GPT-5 data were obtained through systematically designed prompts during late September and early

October 2025, whereas the comparative dataset based on GPT-4o, as stated in Balenović and Proroković (2025), was collected between February and March 2025.

3 Results

3.1 Qualitative Overview of the Lexicographic Data on oanwords in L2 context

When examining the GPT-5 responses to the prompt requesting the senses of English loanwords in Croatian, several notable inconsistencies and interpretive issues emerge (Table 2). For instance, in the case of the loanword *cross*, the model displays considerable inter-response stochasticity—alternating between orthographically adapted and non-adapted forms, offering English examples accompanied by Croatian translations, and at times generating seemingly unattested or fabricated Croatian examples that appear to be semantic overgeneralizations from English (e.g., *Nemoj* ***biti cross***; a supposedly Croatian counterpart of "Don't be cross"). A similar phenomenon is observed with the loanword *touch*, where the derived verbal representations such as *Nemoj to* ***touchati*** or ***Touchaj*** *ikonu da se otvori* cannot be verified through standard web or corpus searches. The lack of attestation suggests that these are not naturalized Croatian usages but rather AI-generated projections extrapolated from English input.

Table 2. Variation across responses and L1 interference in GPT-5 responses to the prompt asking for the senses of the English loanwords *round, dog, cross* and *touch*, in the Croatian language[1]

GPT-5 feedback	Comment
Riječ „round" (izgovor: raund) ušla je u hrvatski jezik kao posuđenica iz engleskog, iako nije jako česta u općem jeziku — ali se koristi u sportu, glazbi, piću i općim izrazima (posebno u žargonu). U nastavku su sva njezina značenja i načini uporabe u hrvatskom jeziku. ***(…)*** *2. „Round" = runda pića (u žargonu) U svakodnevnom govoru, pogotovo među mlađima ili u kontekstu ugostiteljstva, round može značiti i krug pića koji netko plaća društvu. Primjeri:* ***„Ja plaćam ovaj round, ti sljedeći."*** ***„Daj još jedan round piva!" (…)*** *3. „Round" = glazbeni kanon / višeglasje (rjeđe) U glazbenom kontekstu, riječ round može značiti pjesmu u kojoj se više glasova pjeva istu melodiju s vremenskim pomakom, kao u engleskom izrazu singing in a round. Primjer:* ***„'Frère Jacques' je tradicionalni round."*** ***(…)***	Unattested L2 loanword usages (likely overgeneralisations from L1 contexts)
Riječ „dog" (izgovor: dog) potječe iz engleskog jezika i znači „pas", ali u hrvatskom jeziku ta riječ nije ušla kao standardna posuđenica — nije dio općeprihvaćenog standardnog jezika. ***(…)*** *2. U žargonu (neformalno, pod utjecajem engleskog). U svakodnevnom govoru, posebno među mlađima koji koriste engleske izraze, dog se ponekad koristi kao nadimak ili metafora (najčešće u internetskom i urbanom slengu). Primjeri:* ***„Ej, dog, kako si?"*** *(= prijateljski izraz, poput „brate", „frende") „On je pravi dog – odan i pouzdan." (= metaforički: vjeran, dobar)* ***(…)*** *5. U prenesenom (simboličkom) značenju U nekim kontekstima (npr. umjetničkim, glazbenim ili metaforičkim), dog se koristi kao simbol: vjernosti i odanosti; grubosti, instinkta, muževnosti; (u negativnom smislu) nižeg statusa ili nečistoće; Primjeri:* ***„U njemu se budi dog."*** *(= životinjski instinkt, borbeni duh)* ***(…)***	

[1] The text listed in the tables represents a reduced version of the GPT-5 responses to the analyzed prompts; i.e., it has been edited to remove symbols and non-essential elements for the sake of brevity and clarity, retaining only information pertinent to this study. Additionally, the responses have been significantly shortened, retaining primarily those instances interesting for the purposes of this study.

Riječ „cross" (izgovor: kros) potječe iz engleskog jezika (i starofrancuskog crois, lat. crux) i u osnovi znači „križ", „preći" ili „ukrižiti", ovisno o kontekstu. ***(…)*** *.1. „Cross" = križ (simbol, znak) Najosnovnije i izvorno značenje riječi cross jest križ, simbol s dvije okomite crte. U tom smislu, cross se često javlja u nazivima ili kao likovni/religijski motiv. Primjeri: „Red Cross" → Crveni križ;* ***„Cross necklace"*** *→* ***ogrlica s križem (…)*** *3. „Cross" u sportu. U sportu, cross ima nekoliko ustaljenih značenja: U atletici i trčanju – cross (ili kros) znači utrka izvan staze, najčešće po neravnom terenu.* ***„Trčim kros svake godine****." „****Školski kros bio je jučer****." (…) 5. „Cross" = ljut, uzrujan (iz engleskog pridjeva) U engleskom cross može značiti i nervozan, ljutit („He was very cross with me"), ali to značenje nije preneseno u hrvatski osim u duhovitim ili stilski imitacijskim kontekstima. Primjer:* ***„Nemoj biti cross****." (šaljivo: „Nemoj se ljutiti.")* ***(…)***
Riječ „touch" (izgovor: tač) potječe iz engleskog jezika i osnovno joj je značenje „dodir". ***(…)*** *Značenje i vrsta riječi ovise o kontekstu — u hrvatskom se najčešće javlja kao imenica ili glagol (u žargonu), ali i kao dio složenica. 1. „Touch" = dodir (fizički kontakt) Najosnovnije značenje — dodir rukom ili drugim dijelom tijela. Primjeri:* ***„He felt the soft touch of her hand."*** *→* ***Osjetio je nježan dodir njezine ruke. „Touch screen"*** *→* ***zaslon na dodir. (…)*** *2. „Touch" kao glagol („dirati", „dodirivati") Izvorno engleski glagol to touch znači dotaknuti, dodirnuti, imati kontakt. U hrvatskom govoru, osobito u žargonu i IT-u, ponekad se koristi u obliku „touchati" / „tačati", ali to nije standardno. Primjeri:* ***„Nemoj to touchati."*** *→ (žargonski) Ne diraj to.* ***„Touchaj ikonu da se otvori." (…)***

A further example of this overextension can be seen with the loanwords *round* and *dog* (Table 2), or even *show*, one of the most frequently borrowed English words in Croatian. When prompted to identify its possible word classes, GPT-5 proposes several options, one of which includes a verb form allegedly used in everyday speech, illustrated by the sentence *Ne moraš mi to* ***showati****, već sam vidio* (eng. "You don't have to show it to me; I've already seen it"). Although such morphological adaptation (*showati*) might appear plausible, corpus and web searches reveal no empirical evidence of its occurrence in authentic Croatian usage. This strongly suggests that the model is hallucinating plausible-sounding but non-existent lexical forms, extending legitimate English derivational or syntactic patterns into Croatian contexts where they do not exist.

An important aspect of prompt design is the introductory phrase *"Ako postoje"* (Eng. *If they exist*), which is used when requesting the possible senses of a particular word in its L2 context, i.e., in Croatian. It is therefore striking that ChatGPT consistently fails to make a clear distinction between English and Croatian when providing these senses. The introductory phrase itself should indicate that the senses in L2 are not necessarily existent, and as such, should relive the pressure of confirmation bias or affirmatively forced output, which the LLMs are designed to do. The model

often continues to list meanings attested in English, rather than restricting its output to the L2 context. This occurs despite the prompt being carefully formulated both semantically and linguistically: the request explicitly asks for the senses of the loanword in Croatian, and the prompt itself is phrased in L2. Nevertheless, the AI frequently lists examples from the source language, provides uses that are inapplicable in Croatian, or offers translations in which the loanword loses its original orthographic or phonological form. In such cases, the output becomes redundant, as it does not provide the information that was explicitly requested. This highlights a persistent limitation of the model: even when the prompt is unambiguous and clearly targeted, GPT occasionally fails to constrain its responses to the intended linguistic context.

A particularly illustrative example of prompt misinterpretation and partial self-correction can be observed in the GPT-5 responses to the loanword *slow* (Table 3). When prompted with the question *"Ako postoje, koja su moguća značenja/načini uporabe riječi 'slow' u hrvatskom jeziku?"* ("If they exist, what are the possible senses of the word *slow* in the Croatian language?"), the model fails to address the intended inquiry regarding the English loanword's usage in Croatian (L2). Instead, it provides examples of the Croatian equivalent *spor*, effectively reverting to source-language translation rather than analyzing the loanword itself. Furthermore, the model introduces irrelevant grammatical information—such as word class categorizations—despite these not being requested and despite the query being posed in a completely new session, suggesting that the model reverts to default definitional behavior when uncertain about the semantic scope of a prompt.

In the subsequent clarification prompt (*"Da, ali mene zanima korištenje engleske riječi 'slow' u njegovom izvornom obliku u hrvatskom jeziku?"* – "Yes, but I am interested in the use of the English word *slow* in its original form in Croatian"), the model explicitly acknowledges its initial misunderstanding and adjusts its approach (Table 3). It now attempts to generate examples of the English loanword used in Croatian contexts. However, its performance remains inconsistent: while some examples correspond to genuinely attested usage, others are invented or unverified upon corpus or web examination (e.g. *živim slow* or *pustili su slow*). This pattern indicates that although GPT-5 is capable of recognizing its prior misalignment and modifying its interpretive frame, it still exhibits semantic overreach; i.e., producing plausible yet unattested examples in an effort to satisfy user intent. Such behavior underscores the model's pragmatic adaptability, but

also its epistemic unreliability in low-frequency, cross-linguistic contexts, where the boundaries between legitimate borrowing and artificial generation remain blurred.

Table 3. Prompt–response misalignment and source-language transfer in GPT-5 lexical feedback

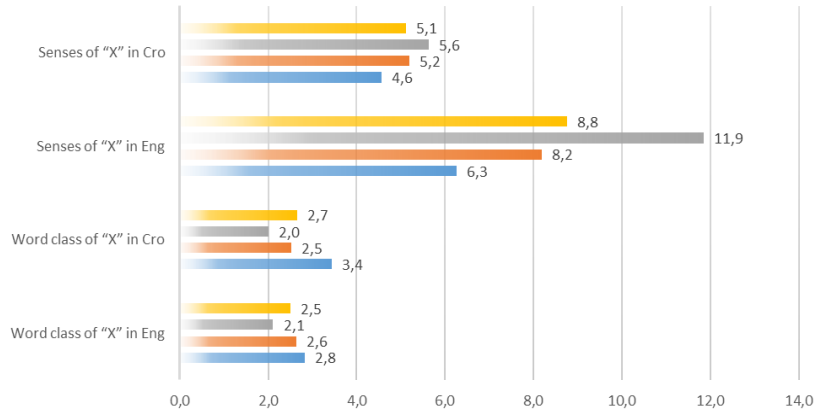
Prompt: Senses of "X" in Cro *Ako postoje, koja su moguća značenja/načini uporabe riječi "slow" u hrvatskom jeziku?*	**Comment**
GPT-5 feedback: *Odlično pitanje! Engleska riječ "slow" u hrvatskom jeziku nema izravni jednoznačni prijevod, ali se može prevoditi i koristiti na više načina, ovisno o kontekstu. Evo glavnih mogućih značenja i uporaba: 1. Kao pridjev – "spor" Najčešće značenje. Osoba, životinja, stvar ili radnja koja se odvija polako. He is a slow runner. → On je spor trkač. The Internet is slow. → Internet je spor. She's slow at understanding jokes. → Spora je u shvaćanju šala. 2. Kao prilog – "polako", "sporo" Kada opisuje način na koji se nešto događa. Drive slow! → Vozite polako! He speaks slow and clearly. → Govori sporo i jasno. (Napomena: u standardnom engleskom bi se reklo "slowly", ali u kolokvijalnom govoru "slow" često zamjenjuje prilog.) 3. Kao glagol (rijetko) – "usporiti" (u izrazu "to slow (down)") U hrvatskom jeziku nema samostalnog glagola slow, nego se uvijek prevodi opisno ili glagolom usporiti. Slow down! → Uspori! The car slowed at the corner. → Auto je usporio na zavoju.* ***(…)***	Providing translated examples of the source language use of the targeted loanword "slow". The model does not give examples for the loanword use in L2, but instead provides examples of how the L1 counterpart for said loanword can be used (*spor*). Additionally, it provides word classes as a response despite not being asked for it, and despite this being a completely new session when the question was posited.
Prompt: Additional clarification for the previous prompt *Da, ali mene zanima korištenje engleske riječi "slow" u njegovom izvornom obliku u hrvatskom jeziku?*	
GPT-5 feedback: *Aha — sad te razumijem točno. Dakle, pitaš ne za prijevode riječi slow, nego na koje se načine sama engleska riječ "slow" (u tom obliku) koristi u hrvatskom jeziku — dakle, kao posuđenica, stilski izraz, oznaka, naziv i sl.* ***(…)*** *Možemo razlikovati nekoliko tipičnih načina uporabe: 1. U nazivima pokreta, trendova i koncepata Najpoznatiji slučaj. Slow food – međunarodni pokret koji promiče lokalnu, tradicionalnu i sporiju pripremu hrane (suprotno "fast foodu"). → U hrvatskom se koristi neprevedeno: pokret slow food, slow food restoran, slow turizam, itd. Primjer: "U Istri raste interes za slow foodom i održivim ugostiteljstvom." Slow travel, slow fashion, slow living, slow journalism itd.* ***(…)*** *2. Kao dio brendova i naziva* ***(…)*** *3. U izražavanju stila života i marketing Koristi se metaforički, često u engleskom obliku i kurzivom: "Trend slow life postaje sve popularniji među mladima." "****Živim slow, bez žurbe****." Tu se slow koristi kao modna riječ, svojevrsni slogan koji implicira filozofiju sporosti, svjesnosti i ravnoteže. 4. U glazbi i umjetnosti Slow jam, slow dance, slow tempo – koriste se u glazbenom žargonu. "****Na kraju večeri pustili su slow****." (u značenju: ples uz sporu glazbu) Ovo je ustaljeno u govoru, osobito među DJ-ima i u plesnom kontekstu.* ***(…)***	The model now acknowledges that it misinterpreted the initial prompt and now attempts to attune to the user's reiterated request. It now demonstrates the understanding of the type of information the user wanted to retrieve, but still fails to provide examples of the loanword truly attested in L2; that is, some of them can be found in everyday use and are confirmed via search engine checks, while others sound odd and fail to withstand this level of scrutiny.

Overall, these examples illustrate GPT-5's continued tendency toward cross-linguistic overgeneralization; that is, its inclination to project L1 semantic and morphological structures onto L2 contexts without sufficient grounding in real language data. This behavior likely stems from the model's training bias toward English-dominant corpora and its design inclination to prioritize affirmative or content-complete responses over epistemic restraint (see Balenović & Proroković, 2025; Malmqvist, 2025). Consequently, while the model demonstrates growing fluency and syntactic adaptability, its semantic precision in bilingual or contact-language contexts remains inconsistent, particularly where corpus evidence is sparse or the prompt demands culturally embedded lexical distinctions.

3.2 Descriptive Statistics: Frequency Data and Trends

The overall results regarding the average number of meanings per prompt are perhaps more revealing in terms of intra-prompt than inter-prompt variation (Figure 1). In other words, it is particularly informative to observe how the number of distinct meanings provided by the model fluctuates in relation to the L2 frequency of a given loanword across different prompts. On one hand, it is unsurprising that the model tends to generate a greater number of senses for a loanword in its L1 context than in its L2 usage, reflecting its broader sense inventories in that language. On the other hand, a notable and somewhat unexpected pattern emerges within both the L1 and L2 "senses" prompts: loanwords that are less frequently attested in L2 usage tend to receive a higher number of meanings from GPT-5 than those that are more frequent. This inverse relationship between frequency and polysemy, as represented in the model's output, runs counter to linguistic expectations. One would expect polysemy to correlate positively with frequency, since more frequent lexical items ought to develop extended senses through broader contextual use. Yet, the model's responses suggest that when an English loanword becomes more frequent in Croatian, the AI attributes fewer potential senses to it. This may indicate that GPT-5 is not sensitive to corpus-based frequency effects in the same way humans are, but rather overgeneralizes from limited or ambiguous cross-linguistic data. Moreover, one might expect the opposite trend in the L1 context if polysemy were indeed a factor influencing borrowing—i.e., that the likelihood of a word being adopted into Croatian increases with its semantic versatility in English. However, this trend does not appear in the results either.

Figure 1. The average number of distinct meanings per prompt type in relation to loanword frequency



	Word class of "X" in Eng	Word class of "X" in Cro	Senses of "X" in Eng	Senses of "X" in Cro
total	2,5	2,7	8,8	5,1
less frequent "X"	2,1	2,0	11,9	5,6
relatively frequent "X"	2,6	2,5	8,2	5,2
highly frequent "X"	2,8	3,4	6,3	4,6

Additionally, the results regarding average numbers of word classes in total are also striking (Figure 1), but they seem to be partly influenced by technically incorrect examples. There is no clear rationale for why a prompt asking for the word classes of a loanword would yield an equal or greater number of categories in its non-native L2 context than in its L1 usage. For instance, consider the example of the word *rock*: although the prompt explicitly requests the word classes of *rock* in Croatian (*if they exist*), GPT produces five different categories (Table 4). Only the first is a legitimate word class (noun); the subsequent categories—such as metaphoric meaning or derived forms—do not correspond to conventional grammatical categories. To be clear, as evident in the example provided in continuation, the model does note in the extension of its (redundant) part of the feedback that it is still referring to specific grammatical category such as nouns, though now in relation to different context use, the question remains as to why it would then separate the number of categories that exist in its L2 use (in comparison to L1 use where the model provided two categories—noun and verb—at the time of the data retrieval, it listed 5 different points)?

Table 4. Feedback inconsistency in word classes of the loanword *rock* in L2 context

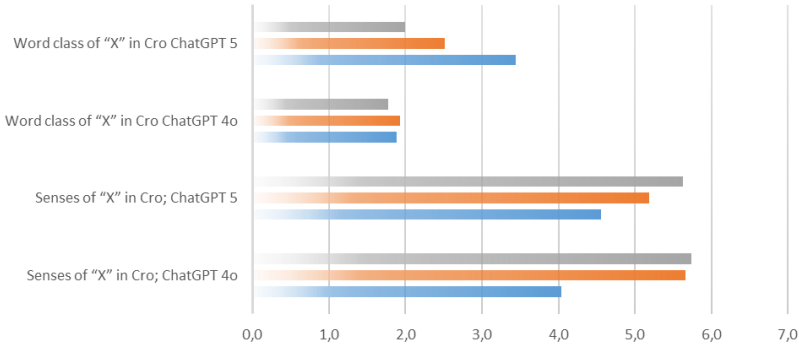
Prompt: Word class of "X" in Cro (*Kao koja vrsta riječi se riječ »rock« može koristiti u hrvatskom jeziku?*)
AI feedback: ***(…)*** *1.* ***Imenica*** *(najčešće i standardno značenje) Vrsta riječi: imenica muškog roda Značenje: glazbeni žanr koji se razvio iz rock'n'rolla; moderna glazba s izraženim ritmom i električnim instrumentima. Primjeri: „Volim slušati rock i blues."* ***(…) 2. Atribut*** */ opisni dodatak (u složenicama)* ***(...)*** *3.* ***Kao sastavni dio naziva ili Brenda*** *Riječ rock često se koristi u nazivima festivala, klubova, pjesama, časopisa, brandova itd. U tim kontekstima zadržava svoj izvorni oblik i piše se velikim početnim slovom ako je dio imena. Primjeri: „INmusic Rock Festival" „Planet Rock Radio"* ***(…)*** *Ovdje je rock imenica, ali funkcionalno dio vlastitog imena.* ***4. Preneseno značenje (kolokvijalno)*** *U neformalnom govoru rock se može koristiti prenesenim značenjem – za nešto što je „jako", „žestoko", „buntovno", „kul". Primjeri: „To je pravi rock!" „Njegov govor bio je čisti rock'n'roll." I dalje je imenica, ali s metaforičkim značenjem.* ***5. Izvedenice*** *(od engleskog korijena). Od riječi rock razvile su se i domaće izvedenice, koje su u potpunosti prilagođene hrvatskom jeziku: Riječ Vrsta riječi Značenje rokerski pridjev koji se odnosi na rock (rokerska glazba, rokerski stil) roker imenica osoba koja sluša ili svira rock* ***(…)***

If the model were uncertain about the intended meaning of the term *"vrsta riječi"* in the prompt—despite it being a well-established term in Croatian language—it would have been preferable for it to seek clarification from the user before generating a response. Instead, the model opts to provide content at the risk of being incorrect or redundant. While this behavior is not inherently a mistake, it highlights a characteristic that is suboptimal for a responsible tutoring or lexicographic system: the generation of intra-prompt variability that exceeds the user's informational needs. Specifically, the model acknowledges the original question but then continues to produce additional categories that are irrelevant or inconsistent, resulting in inadequate outputs. For example, if asked in English, it would be illogical to classify a word simultaneously as a noun, a verb, and a metaphoric or derived unit. This pattern again underscores both the stochastic nature of LLM outputs and the importance of human oversight: users must critically evaluate responses and, ideally, the model should be designed to seek clarification when prompts are ambiguous or when generating outputs that could be misleading.

The results obtained with the GPT-5 model reveal patterns similar to those observed in the earlier study by Balenović and Proroković (2025), which was based on the GPT-4o dataset used for comparison in Figure 2. In both cases, the same inverse relationship emerges between the frequency of a loanword in Croatian and the number of meanings that the model attributes to it: words occurring less frequently in the L2 context are typically associated with a greater number of senses. On the other hand, the number

of different word classes suggested by the AI models either rise with the frequency of the loanwords or remain stable.

Figure 2. Comparing GPT4o and GPT5 outputs in relation to loanword frequency and the number of meanings per prompt



	Senses of "X" in Cro; ChatGPT 4o	Senses of "X" in Cro; ChatGPT 5	Word class of "X" in Cro ChatGPT 4o	Word class of "X" in Cro ChatGPT 5
less frequent	5,7	5,6	1,8	2,0
relatively frequent	5,7	5,2	1,9	2,5
highly frequent	4,0	4,6	1,9	3,4

As far as the number of reverse proportionality in the sense-frequency relationship, Balenović and Proroković argued that one plausible explanation for this trend was in the model's access to data. For more frequent items, the abundance of contextual evidence enables the model to delimit meanings more precisely, resulting in fewer but more accurate senses. Conversely, for low-frequency loanwords, the scarcity of L2 examples leads the model to overgeneralize, often transferring semantic distinctions from English into Croatian contexts where they are not attested. While the GPT-5 model demonstrates moderate improvement over its predecessor in managing these distinctions, the same tendency persists: the model continues to exhibit semantic inflation for less commonly used loanwords (cf. Table 1). This suggests that the problem is not merely one of model architecture, but of data distribution and representativeness. As the previous authors note, such outcomes may also be affected by the corpus compilation methodology (in this case, the ENGRI corpus) which, being domain-specific, could skew frequency counts and thereby influence interpretive results. Further comparative testing on non-domain-specific datasets could therefore help determine whether these trends are intrinsic to the model's processing or contingent upon corpus design.

3.3 Inferential Statistics: Frequency Effects on Polysemy and Word Class Variation

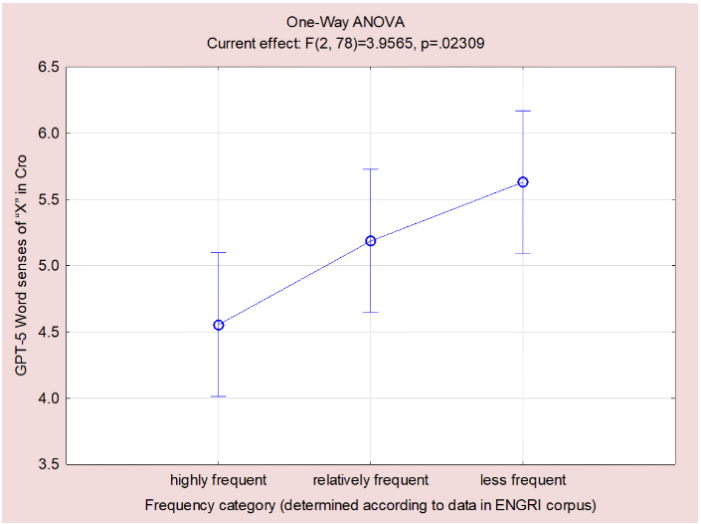
A one-way ANOVA was conducted to examine the effect of loanword frequency in the L2 (Croatian) on two dependent variables: (1) the number of word senses that the GPT-5 model generated for each item in Croatian, and (2) the number of distinct word classes assigned to those items. The analysis revealed a statistically significant effect of frequency category on the number of word senses, $F(2, 78) = 3.9565$, $p = .02309$. As shown in Figure 3, the mean number of senses increased as frequency decreased, suggesting that less frequent loanwords tended to exhibit a greater range of semantic interpretations in Croatian. In other words, with the decline in the frequency of a loanword in L2, there is a rise in the number of senses it exhibits in GPT-5's output. Conversely, the results for word classes indicated the opposite trend. The one-way ANOVA for word class variation yielded a significant main effect of frequency, $F(2, 78) = 15.791$, $p < .001$. As displayed in Figure 3, the number of distinct word classes generated by the model decreased with lower-frequency items. This inverse relationship suggests that while GPT-5 tends to overgeneralize semantically when examples are sparse, it remains more constrained grammatically under similar conditions.

This pattern aligns with the descriptive findings and further supports the interpretation that, for low-frequency loanwords, the scarcity of L2 evidence in the input data may lead the model to transfer semantic distinctions from English into Croatian contexts where such senses are not attested. On the other hand, the more stable and proportionally narrower distribution of word classes may indicate that the model, when not constrained by prompts demanding elaborate output (the implication being that the ask for the number of possible word classes is inherently one less demanding), achieves greater accuracy. This asymmetry highlights the complex relationship between input frequency, lexical representation, and model behaviour.

The correlation matrix presented in Table 5 examines the relationship between GPT-4o and GPT-5 outputs for the number of word senses and word classes generated for English loanwords in Croatian, alongside their frequency in the L2. The analysis reveals a significant positive correlation between GPT-4o and GPT-5 in the number of senses assigned to the same items ($r = .42$, $p < .01$), indicating a consistent trend across models in how they semantically interpret loanwords in Croatian. In other words, when GPT-4o generated a greater number of senses for a given word, GPT-5 tended to do

the same, suggesting a shared underlying representational bias or overlap in their training data and lexical mapping.

Figure 3. One-Way ANOVA analysis for frequency-polysemy relationship in GPT-5 model when it comes to both word senses and word classes



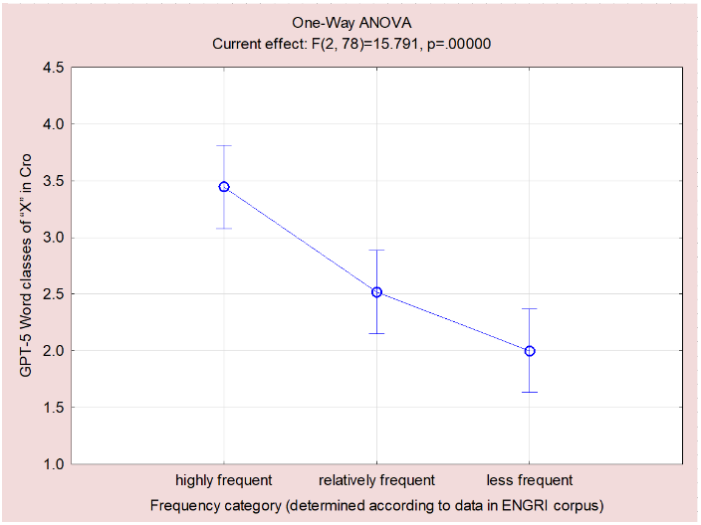


Table 5. Correlation matrix concerning output sense frequency trends across the two examined models (GPT-4o and GPT-5)

	GPT-4o Senses of "X" in Cro	GPT-4o Word class of "X" in Cro	GPT-5 Senses of "X" in Cro	GPT-5 Word class of "X" in Cro	Loanword frequency in Croatian
GPT-4o Senses of "X" in Cro	1.00	0.11	0.42**	-0.01	-0.25*
GPT-4o Word class of "X" in Cro		1.00	0.22	0.14	0.09
GPT-5 Senses of "X" in Cro			1.00	-0.00	-0.09
GPT-5 Word class of "X" in Cro				1.00	0.28*
Loanword frequency in Croatian					1.00

*p<0.05; **p<0.01

In contrast, the correlation between word class variation and frequency was positive but modest (r = .28, p < .05 for GPT-5), indicating that more frequent loanwords are associated with a greater diversity of syntactic realizations, a result that complements the ANOVA findings. No significant cross-model correlation was observed for word classes, implying that GPT-5 and GPT-4o diverge more in grammatical than in semantic mapping when producing Croatian output. Taken together, these results suggest that both models exhibit a consistent pattern in their semantic behavior, but GPT-5's outputs remain more stable and contextually constrained in terms of grammatical categorization, thus potentially supporting the view that model evolution between GPT-4o and GPT-5 primarily refines morphological and syntactic inference, while maintaining a similar approach to lexical-semantic generalization though this inference remains beyond the aims and scope of this paper and analysis.

4 Discussion

Out of the issues recently compiled by the European Data Protection Board in *AI Privacy Risks & Mitigations—Large Language Models (LLMs)* (Barberá, 2025), including context-related knowledge limitations, lack of robustness,

and variability in training data quality, the challenge most relevant to this study concerns prompt design and input quality. As noted in the report, LLMs are highly sensitive to the phrasing of inputs, with minor variations in prompt formulation often leading to drastically different outputs (pp. 22–23). In the context of this research, it is useful to highlight a related phenomenon that is not explicitly underscored in Barberá (2025): even when identical prompts are submitted in separate, newly initiated sessions, the model frequently produces different outputs. In this study, the only variable affecting these inter-prompt differences is the specific loanword being analyzed. This observation reinforces the stochastic nature of LLM responses: output variability is not solely a function of prompt formulation but also of inherent model randomness. The model may offer diverse avenues for defining the same loanword in an L2 context, reflecting a degree of unpredictability that must be taken into account in lexicographic applications.

Furthermore, the educational implications of this lexicographic research are highly relevant in the context of integrating AI into language learning, and they seem to go far beyond the quality of the model's output; they concern the quality of the input—that is, what and how users ask. Even when an AI system is technically accurate and contextually fluent, its performance remains bounded by the specificity and appropriateness of the prompt it receives. A poorly designed prompt can easily yield an answer that is formally correct yet functionally misleading: it may fail to provide the information the user actually intended to retrieve. In this sense, the "failure" of the AI lies precisely in its success; it performs its task with perfect obedience to the input, yet lacks the inferential, social, and contextual awareness that a human interlocutor would naturally apply to interpret the learner's intent. These implications align well with recent findings indicating that users of AI models have to develop the metacognitive ability to evaluate whether the data retrieved truly satisfies the learning task at hand (Woo et al., 2024; Kang et al., 2025; Kim et al., 2025), suggesting that human–AI co-agency in language learning is one of the ways to mediate the negative effects (cf. Alm, 2025; Ciocan, 2025). Ultimately, as Liu, Chen, and Xu (2025) demonstrate, AI tools are not (yet) sufficient to simply replace traditional learning instruments, but instead are to reshape the learner's relationship when it comes to knowledge retrieval and evaluation, making reflective engagement and guided mediation more crucial than ever.

Unlike human teachers, who rely on pragmatic and paralinguistic cues to infer a learner's communicative goals, AI systems cannot read beyond the literal formulation of the query. They respond to "what is asked", not

necessarily to “what is meant”. This asymmetry highlights a critical limit of learner autonomy in AI-mediated learning environments: effective interaction with AI presupposes a high level of prompt literacy: the ability to formulate questions precisely, interpret responses critically, and assess whether the output aligns with one’s learning objectives. Without such literacy, learners risk misinterpreting or over-trusting outputs that, while linguistically plausible, may not serve their intended purpose. Moreover, this issue may be even more relevant in the context of response accuracy; i.e., user’s lack of critical reflection is not only problematic in relation to ChatGPT's responses failure to address the intended purpose of the prompt, but further amplifies with the possibility of said AI model to provide accurate or truthful answers. This is another issue identified by Barberá (2025), which she labels as inadequate calibration, resulting in model overconfidence. This study firmly corroborates that observation: the AI often provides unattested examples of loanwords in Croatian with high confidence. Correcting such errors requires an interlocutor who not only recognizes the mistake but is also capable of critically evaluating the AI’s output. Consequently, effective interaction with LLMs demands that human users possess both sufficient linguistic competence and a healthy degree of skepticism, enabling them to challenge and refine the model’s responses (cf. Spatola, 2024).

Additionally, the results also clearly outlined the inverse relationship between frequency and polysemy which, as represented in the model’s output, runs counter to linguistic expectations. Typically, corpus-based studies demonstrate that polysemy correlates positively with frequency, since more frequent lexical items tend to develop extended senses through broader contextual use (see Zipf, 1949; Ferrer-i-Cancho, 2014; Casas et al., 2019; Ferrer-i-Cancho & Vitevitch, 2018 and others). Yet, the model’s responses suggest that when an English loanword becomes more frequent in Croatian, the AI attributes fewer potential senses to it. Moreover, one might expect the opposite trend in the L1 context if polysemy were indeed a factor influencing borrowing; i.e., that the likelihood of a word being adopted into Croatian increases with its semantic versatility in English (cf. Chesley & Baayen, 2010; Ingham, 2024). However, this correlation does not appear in the results either (cf. Calude et al., 2020). Together, these findings suggest that while GPT-5 demonstrates considerable lexical sensitivity, it does not model frequency-based semantic extensions in a human-like manner, and thus fails to replicate expected relationships between loanword frequency, semantic range, and contextual adaptability.

5 Conclusion

This study underscores the growing importance of cross-linguistic lexicographic research that integrates artificial intelligence tools. This type of research can help us, not only advance our understanding of how large language models process meaning and context across languages, but also understand broader pedagogical implications for how AI is to be used in educational settings. In this sense, the findings presented here highlight two major observations:

(1) In the educational and lexicographic context, the most important implication of this research lies in highlighting the user's responsibility to approach AI-generated linguistic data critically. While large language models such as GPT-5 can provide comprehensive and often insightful lexical analyses, they also display a non-negligible degree of inaccuracy and contextual misalignment. Consequently, users who are insufficiently aware of these limitations (or lack the requisite competence to assess them), risk accepting flawed or misleading interpretations as accurate. This underlines the need to develop not only linguistic and lexicographic literacy but also AI literacy: the ability to design precise prompts, to recognize the boundaries of model reliability, and to maintain a healthy degree of skepticism when interpreting outputs.

(2) The results further confirm the observations made by Balenović and Proroković (2025), indicating that GPT-based models, despite notable advances, still exhibit inconsistency in distinguishing between the senses of English loanwords used in Croatian (L2) contexts. The findings suggest that the less frequent a loanword is in everyday Croatian use, the higher the likelihood of the model overgeneralizing from its English (L1) counterpart; i.e., often extending senses that are not attested in the recipient language. Moreover, the comparative analysis between GPT-5 and GPT-4o reveals similar tendencies, with counterintuitive results showing a greater number of proposed senses for high-frequency items. This pattern appears to reflect ChatGPT's inclination to prioritize response generation over precision, where overgeneralization and L1-to-L2 semantic transfer contribute to inflated sense counts in less frequent or semantically ambiguous categories.

References

Alm, A. (2025). Reconceptualising literacy as co-literacy in language education with AI. *Insights into AI and Language Teaching and Learning*. Castledown Publishers, London. https://doi.org/10.29140/9781763711600-04

Atkins, B. S., & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Balenović, K., & Proroković, J. (2025). The lexicographic potential of artificial intelligence: a case study of English loanwords in the Croatian language. *Studia lexicographica: časopis za leksikografiju i enciklopedistiku*, *36*(19), 39-64. https://doi.org/10.33604/10.33604/sl.19.36.3

Barberá, I. (2025). *AI Privacy Risks & Mitigations—Large Language Models (LLMs). European Data Protection Board.* Available online: https://www.edpb.europa.eu/system/files/2025-04/ai-privacy-risks-and-mitigations-in-llms.pdf (accessed on 12 June 2025).

Beliga, S., & Filipović Petrović, I. (2024). Large language models supporting lexicography: conceptual organization of Croatian idioms. In *Conference on Language Technologies and Digital Humanities, Ljubljana 2024*, 23–45.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big? 🦜 In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency,* 610-623. https://doi.org/10.1145/3442188.3445922

Bergenholtz, H., & Tarp, S. (2003). Two opposing theories: On H.E. Wiegand's recent discovery of lexicographic functions. *Hermes*, *31*, 171–196.

Bogunović, I. & Kučić, M. (2022). *The Database of English Words in Croatian.xlsx. figshare. Dataset.* https://doi.org/10.6084/m9.figshare.20014364.v1.

Bogunović, I. (2023). A corpus-based approach to English loanwords: Introducing the database of English loanwords in Croatian. *Fluminensia*, 35 (2): 437–60. https://doi.org/10.31820/f.35.2.1

Calude, A., Miller, S. & Pagel, M. (2020). Modelling loanword success – a sociolinguistic quantitative study of Māori loanwords in New Zealand English . *Corpus Linguistics and Linguistic Theory*, *16*(1), 29-66. https://doi.org/10.1515/cllt-2017-0010

Casas, B., Hernández-Fernández, A., Català, N., Ferrer-i-Cancho, R., & Baixeries, J. (2019). Polysemy and brevity versus frequency in language. *Computer Speech & Language*, *58*, 19-50. https://doi.org/10.1016/j.csl.2019.03.007

Chen, L., Dao, H. L., & Do-Hurinville, D. T. (2024, September). AI empowerment: where are we in the automation of lexicography? A metaphraseographic study. In *ASIALEX 2024-The 17th International Asian Association for Lexicography Conference*.

Chesley, P., & Baayen, R. H. (2010). Predicting new words from newer words: Lexical borrowings in French. *Linguistics*, *48*(6). https://doi.org/10.1515/LING.2010.043

Ciocan, M. (2025). Integrating AI without losing the human element in the English language classroom. *Dialogo*, *11*(2), 71-80.

de Schryver, G. M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, *36*(4), 355-387. https://doi.org/10.1093/ijl/ecad021

Ethayarajh, K. (2019). How contextual are contextualized word representations? Comparing the geometry of BERT, ELMo, and GPT-2 embeddings. *arXiv preprint arXiv:1909.00512*. https://doi.org/10.48550/arXiv.1909.00512

Ferrer-i-Cancho, R. (2014). The meaning-frequency law in Zipfian optimization models of communication. *arXiv preprint arXiv:1409.7275*. https://doi.org/10.48550/arXiv.1409.7275

Ferrer-i-Cancho, R., & Vitevitch, M. S. (2018). The origins of Zipf's meaning-frequency law. *Journal of the Association for Information Science and Technology*, *69*(11), 1369-1379. https://doi.org/10.1002/asi.24057

Ingham, R. (2024). Loanwords and polysemy: An investigation of specialized domain lexis in Middle English. *Lexis. Journal in English Lexicology*, (HS 3). https://doi.org/10.4000/12izc

Kang, X., Li, X., Bai, X., Zhang, Y., Chen, S., et al. (2025). Exploring the structural logic and learning path of prompt language in AI-assisted interaction design. *Education Journal*, *14*(3), 126-133. https://doi.org/10.11648/j.edu.20251403.15

Kim, J., Yu, S., Lee, S. S., & Detrick, R. (2025). Students' prompt patterns and its effects in AI-assisted academic writing: Focusing on students' level of AI literacy. *Journal of Research on Technology in Education*, 1–18. https://doi.org/10.1080/15391523.2025.2456043

Klosa-Kückelhaus, A., & Tiberius, C. (2025). The lexicographic process revisited. *International Journal of Lexicography*, *38*(1), 1-12.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, *11*(1), 1-8. https://doi.org/10.1057/s41599-024-02889-7

Liu, R., Chen, X., & Xu, Y. (2025). Beyond replacement: how large language models influence dictionary usage patterns among Chinese English learners. *International Journal of Lexicography*, ecaf017. https://doi.org10.1093/ijl/ecaf017

Liu, J., Shen, D., Zhang, Y., Dolan, B., Carin, L., & Chen, W. (2021). What makes good in-context examples for GPT-3?. *arXiv preprint arXiv:2101.06804*. https://doi.org/10.48550/arXiv.2101.06804

Malmqvist, L. (2025). Sycophancy in large language models: Causes and mitigations. In *Intelligent Computing-Proceedings of the Computing Conference* (pp. 61-74). Cham: Springer Nature Switzerland.

Milakis, E. D., Argyrakou, C. C., Melidis, A., & Vrettaros, J. (2025). ChatGPT and AI chatbots in education: An umbrella review of systematic reviews, scoping reviews, and meta-analyses. *International Journal of Education and Information Technologies*, *19*, 100-119. https://doi.org/10.46300/9109.2025.19.11

Ptasznik, B., & Lew, R. (2025). Dictionaries versus AI tools through the eyes of English majors. *International Journal of Lexicography*, *38*(2), 140-158. https://doi.org/10.1093/ijl/ecaf005

Spatola, N. (2024). The efficiency-accountability tradeoff in AI integration: Effects on human performance and over-reliance. *Computers in Human Behavior: Artificial Humans*, *2*(2), 100099. https://doi.org/10.1016/j.chbah.2024.100099

Woo, D. J., Wang, D., Yung, T., & Guo, K. (2024). Effects of a prompt engineering intervention on undergraduate students' AI self-efficacy, AI knowledge and prompt engineering ability: A mixed methods study. *arXiv preprint arXiv:2408.07302*. https://doi.org/10.48550/arXiv.2408.07302

Yarbro, J. T., & Olney, A. M. (2021, June). Contextual definition generation. In *Proceedings of the Third International Workshop on Intelligent Textbooks* (Vol. 2895).

Zipf, George K. (1949). *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.

Abstrakt: Tento výzkum replikuje experimentální design Balenoviće a Prorokoviće (2025) na stejném lexikálním datovém souboru a zkoumá, zda model GPT-5 poskytuje přesnější, kontextově vhodnější a jazykově diferencovanější lexikografický výstup, pokud jde o významy a použití anglických přejatých slov v chorvatštině. Cílem je dále zjistit, jak se definice, kontextové příklady a rozdělení významů modelu GPT-5 kvalitativně a kvantitativně liší od těch, které produkuje model GPT-4o. Navzdory zlepšením výsledky ukazují, že model GPT-5 nadále vykazuje systematické slabiny v kontextu L2, zejména u málo frekventovaných přejatých slov. Ačkoli ne výlučně, stejně jako v GPT-4o, méně doložené položky vyvolávají vyšší míru nadměrné generalizace z L1 do L2, což vede k neověřeným nebo sémanticky nepravděpodobným použitím prezentovaným s vysokou mírou jistoty. Studie dospěla k závěru, že ačkoli GPT-5 představuje významný pokrok v lexikografické oblasti, jeho výkonnost zůstává při interpretaci přejatých slov v L2 nejednotná. Jinými slovy, stále se zdá, že existuje potřeba kritického lidského dohledu a vzdělávacího významu gramotnosti v oblasti AI. Nesprávná interpretace návrhu a záměru modelu ze strany modelu znamená, že rozpoznání a oprava takových chyb vyžaduje nejen povědomí uživatele, ale také jazykovou kompetenci.

Klíčová slova: velké jazykové modely (LLM), GPT-5, mezijazykové lexikografie, definice významu přejatých slov, nadměrná generalizace a přehnaná sebejistota umělé inteligence

Dr. Jakov Prorokoviće,
Postdoctoral Fellow at the Department of Teacher and Preschool Teacher Education
University of Zadar
jprorokov@unizd.hr