

So Close but Still Far: Case Study on Application of LLMs in Idioms Identification, Definition, and Generation of Illustrative Examples

Aleksandra Marković¹, Ranka Stanković²

¹Institute for the Serbian Language SASA, Belgrade, Serbia

² University of Belgrade, Faculty of Mining and Geology, Belgrade, Serbia
E-mail: aleksandra.markovic@isj.sanu.ac.rs, ranka.stankovic@rgf.bg.ac.rs

Abstract

Automation has revolutionised lexicography, introducing the ‘post-editing lexicography’ model, where the role of the lexicographer involves refining automatically generated dictionary drafts. Since the launch of ChatGPT in November 2022, numerous papers have explored the potential applications of LLMs in dictionary production. The rapid evolution of LLMs necessitates a re-evaluation of conclusions drawn approximately two years prior regarding their application in automating dictionary entry creation, particularly in light of the advanced capabilities demonstrated by contemporary models.

We will illustrate an experiment conducted on a dataset of 400 (397) MWEs with idiomatic meaning, aiming to evaluate the usefulness of LLMs in Serbian descriptive lexicography tasks (idiom generation, word-sense disambiguation of MWEs, definition writing, and generation of illustrative examples). We requested two types of illustrative examples: those in which a MWE has an idiomatic meaning, and examples with that meaning paraphrased literally (without the idiom). We will highlight the challenges and issues encountered with several models (ChatGPT-4o and 4.1, Gemini-2.5-Flash and 2.5-Pro) and discuss the differences in their performance based on given LLM prompts using direct chat and APIs access via Python scripts.

Keywords: descriptive monolingual lexicography; LLMs; word sense disambiguation; generating idioms, definitions and illustrative examples; Serbian language

1. Introduction

The application of LLMs in lexicography raised the question of whether a human lexicographer is still needed. The answer is still positive. Automation has revolutionised lexicography, introducing the “post-editing lexicography” model, where the lexicographer’s role involves refining automatically generated dictionary drafts (Rundell, 2024).

Since the launch of ChatGPT in November 2022, numerous papers have explored the potential applications of LLMs in dictionary production. Some authors express optimism (de Schryver & Joffe, 2023), while others are somewhat sceptical or cautious (Jakubiček & Rundell, 2023). They base their opinion on experiments in which ChatGPT was used to create entry components or complete entries for a mini monolingual English dictionary (or even for a small bilingual dictionary, like (de Schryver & Joffe, 2023)). ChatGPT struggled with word sense disambiguation (WSD) and with generating illustrative examples, while the definitions were assessed as “well-written and accessible” (Rundell, 2024). Phoodai

et al. (2023) demonstrated that ChatGPT performed lower than the Oxford Advanced Learner’s Dictionary (OALD) in terms of MWE.

We will illustrate an experiment conducted on a dataset of 400 (397 more precisely) MWEs with idiomatic meaning, aiming to evaluate the usefulness of LLMs in Serbian descriptive lexicography tasks, specifically word sense disambiguation of MWEs, definition writing, and generation of illustrative examples. We prompted for two kinds of illustrative examples: one in which a MWE has an idiomatic meaning (sr idiom *preko trnja do zvezda*, en *through hardship to the stars*), and an example with that meaning expressed by different wording (not the idiom). For example, the model offered MWE’s definition (en ‘A way to success leads through pain and difficulties’) plus two illustrative examples, one with this idiom (en ‘His way through life was difficult, an excellent example [of] through hardship to the stars’) (this example in sr wasn’t well-formed), and the other with this idiomatic meaning, but without the idiom itself (en ‘His way through life was difficult, he succeeded overcoming pain and difficulties’), and this example in sr was assessed as good (for evaluation criteria see 4). To make this task even more complex, we also requested the models to generate Serbian language idioms and perform all the requested tasks on them. We will highlight the challenges and issues encountered with several models (ChatGPT-4.0 and 4.1, Gemini-2.5-Flash and 2.5-Pro) and discuss the differences in their performance based on given LLM prompts using direct chat and APIs access via Python scripts.

The first evaluation results from our pilot study, gained from a subset of 50 MWEs generated by Gemini-2.5-Pro, showed that it performs these tasks reasonably well. There are still some challenges: lemmatisation issues; WSD (sometimes, when a MWE has more than one sense, a model generates a definition that combines both meanings); definitions were rather good, but they, as well as the illustrative example, needed post-editing.

The rest of the paper will be organised in five sections. Section 2 will introduce related work. Section 3 introduces the aim of the research. In section 4, we describe the three experiments we conducted, and the methodology we applied in the assessment of the results obtained. Results and discussion follow in 6, and then concluding remarks in 7.

2. Related Work

As mentioned, since its launch at the end of 2022, numerous papers have explored the results of using LLMs in lexicography. The most up-to-date lexicography model in the second decade of the XXI century was, according to Rundell (2024), the post-editing model, in which machines can take over many dictionary-making tasks – the appropriate software chooses the relevant information from a corpus and fills in the dictionary database. The lexicographer’s job is to evaluate the first draft of the dictionary entry produced by the machine and to decide on what to keep or delete, or what should be added. Relevant to our work is the observation that, in this model, one task was dependent on human labour, and “refused” to be automated, namely, definition writing Rundell (2024).

Since ChatGPT appeared in 2022, and the other LLMs followed, this author analysed the post-editing model in opposition to the model that can produce almost a finished dictionary in one step. This analysis aimed to address three major questions: Do we still need dictionaries, since we can pose a question about meaning directly to the LLM? Do we still need lexicographers, given the availability of LLMs? And do we still need tools

designed for various dictionary-making tasks Rundell (2024)? Happily, the answers to all these questions are positive. Good reference dictionaries offer reliable information; lexicographers are still irreplaceable, since models provide only a dictionary draft; and last, existing tools display better results¹. Generally, the problem with LLMs is that we do not know what they base their responses on, so what is missed in this model is the vital interaction with the corpus.

Jakubíček & Rundell (2023) experimented with the automatic production of a small English language dictionary, consisting of 99 simple entries, using ChatGPT (3.5 and 4). The paper highlights the advantages and disadvantages of utilising LLMs in lexicographic tasks, and the results provided by ChatGPT are compared with those of existing tools and technologies. As for the lexicographic evaluation, what was assessed was meaning identification, the quality of definitions, grammatical information, of labelling marked lexis, and illustrative examples. The output was compared with the information offered by reference English language dictionaries.

Shortly, there was a problem with meaning identification, as models sometimes present *false polysemy* (identifying non-existent meanings) or present the same meaning with different definitions. This happens even in simple polysemic structures, chosen for the experiment (which implies that more complex polysemy would be even harder to tackle). One of the important conclusions is that the non-deterministic nature of the model makes it hard to evaluate even the results of a single prompt. As for the definitions, the authors reckon that their writing is one of ChatGPT’s stronger points, since it uses the formulae of existing dictionary definitions on which it was trained.² Marked lexical items were mostly properly labelled. Among the advantages are also the affordability, the fact that prompts could be refined, and that the models will be continuously developed. What is clear, though, is that the lexicographer will still be needed in the future.

Lew (2023) also presented an experiment on using ChatGPT in lexicography. He designed his experiment choosing 15 verbs of communication (since those verbs have specific complementation types, as well as pragmatic conditions of use). The author provided the prompt with two dictionary entries from the COBUILD Online dictionary (for the verbs “approve” and “assassinate”, the latter not being a communication verb), serving as a template for generating dictionary entries for 15 chosen verbs. Four experienced lexicographers evaluated the results, and their feedback was used to refine the prompts further. The conclusion was that ChatGPT (Plus model) was capable of writing full-sentence definitions, typical of COBUILD dictionaries, and that these definitions were difficult to differentiate from those written by experienced and trained lexicographers. As for illustrative examples, Lew found out that their quality wasn’t as good, but it could be refined by prompt engineering.

de Schryver (2024) emphasizes that the most successful experiments on LLMs in lexicography were conducted for the English language. The first attempt at applying ChatGPT to

¹ Rundell illustrates this claim with the well-known example of the verb *to cause* – ChatGPT, when prompted to explain meaning and use of this verb, managed to “give a well-written, discursive response, but repeatedly fell short of an adequate explanation”. By this, the author wishes to tell that the phenomenon of negative prosody, easily noticeable through Word Sketches, couldn’t be noticed by ChatGPT.

² But the authors observe that definitions sometimes miss mentioning key meaning components, like when defining *garden*, to mention the fact that it is typically attached to a house.

the Portuguese language was made by the author himself, who noticed that the results were poor because the model was trained on English data and for English, and that these results were then translated into Portuguese. Also, the model was not adapted to some exotic languages. De Schryver says that the possibility of adapting ChatGPT for lexicographic tasks made a difference.

Filipović-Petrović & Beliga (2024) reports on the experiment in which ChatGPT-4 was asked to differentiate the examples of use of the hr idiom *isplivati na površinu* in which the idiom had metaphorical meaning, from those with literal meaning. The model correctly identified 248 examples and incorrectly identified 132. The authors conclude that the model should be able to complete tasks like this without errors, to be useful for lexicography purposes.

3. A small dictionary of Serbian idioms generated by LLMs

This research aims to analyse the possibilities of applying LLMs in Serbian as an under-resourced language, specifically in idioms generation and identification, writing definitions, and generating illustrative examples for the Serbian Language. With that purpose in mind, we created a small dictionary of Serbian idioms. We requested the model to generate (the total of) 400 idioms in the Serbian language, to perform WSD (if there was more than one sense), than to generate the definition(s) (for each identified sense) in standard Serbian, as well as two kinds of illustrative examples, one with the idiom, and the other without the idiom, but with an idiom’s meaning preserved in the paraphrased sentence example.

The rationale behind prompting for an example with an idiom’s meaning, but without the idiom, lies in the fact that, as (Gantar, 2024) points out, results of such prompts allow us to evaluate the level of LLMs’s “understanding” of idioms. In addition, one of the author’s conclusions is that ChatGPT often provides a literal interpretation of an idiom that it does not understand or misunderstands. The author believes that if we aim to teach a model human language, it makes sense to provide the training set with both examples that convey idiomatic and literal meanings. (Filipović-Petrović & Beliga, 2024) report on different aspects of the importance of idioms’ semantics (linking phraseological synonyms, the need for separating the instances of literal and idiomatic usage of idioms for specific tasks), and (Beliga & Filipović Petrović, 2024) report on categorising idioms into semantic fields.

It is worth mentioning that we are engaged in a shared task in the UniDive COST Action, namely PARSEME 2.0 Multilingual Shared Task on Identification and Paraphrasing of Multiword Expressions³.

4. Experiments

We decided to evaluate the results along the following axes:

- A1. Is this the idiom of the Serbian language? To answer this question, we consulted reference descriptive (Gortan Premk, D. et al., 1959–2023; Stevanović, M. et al., 1982) and phraseological (Kovačević, 2002; Otašević, 2012) dictionaries of the Serbian

³ <https://unidive.lisn.upsaclay.fr/doku.php?id=other-events:parseme-st>

language, as well as our competence (for rare cases when an idiom is well known and frequent, but not represented in any dictionary).

- A2. Is there more than one sense identified? To assess the polysemic structure (if represented by the model), we also used the dictionaries mentioned above.
- A3. Is the definition (for each sense, if there is more than one):
- acceptable (does it appropriately describe the meaning of the idiom; written using standard Serbian language, without grammatical errors, misspellings, etc.; does the definition correspond to the idiom’s part-of-speech (e.g., if the idiom has an adverbial meaning, is the definition following that)?
 - unacceptable?
- A4. Is the illustrative example with an idiom:
- acceptable (does it represent the idiom’s usage appropriately; is it following the Serbian grammar; does it sound natural (for example, does it contain frequent collocations)?
 - unacceptable?

The assessment of illustrative examples, in addition to the criteria we mentioned, was aided by the criteria for good dictionary examples (Kilgarriff et al., 2008; Stanković et al., 2019).

- A5. Is the example paraphrasing the idiom’s meaning understandable and acceptable? For the assessment of the paraphrase, we used our lexicographic and linguistic expertise.

For the current research, we didn’t have the inter-annotator agreement statistics, and we are aware of this shortcoming; nevertheless, our own lexicographic competence, combined with the reference dictionaries we consulted, makes the evaluation process less subjective.

This piece of research builds on the set of three experiments on the application of LLMs in the Serbian descriptive lexicography (Marković & Stanković, 2025). In this paper, we present a set of evaluation criteria for lexicographic definitions. These criteria involve assessing formal and substantive aspects.⁴ The nature of the experiments we present in the current paper deals with idioms; that is why we couldn’t apply the same, more objective evaluation criteria for definitions.

4.1 Experiment 1

The first experiment was conducted with the Gemini- 2.5-Pro and Gemini-2.5-Flash, with the following prompt (in the Serbian language):

You are a lexicographer and you have to generate 50 idioms in the Serbian language for which you need to give a descriptive definition for each meaning, and for each meaning, you need to give one example, in which the idiom will be used, and another example, with the meaning of the given idiom but not the idiom itself.

The generated text should be written in Cyrillic, standard Serbian, and structured in TSV (Tab-Separated Values) format, with the fields: id, mwe, definition, example1, example2.

For example:

```
imati čdugaak jezik (TAB) šprevie čpriati, biti brbljiv, ogovarati (TAB) Pazi šta čšpria pred njom, ima čdugaak jezik (TAB) Pazi šta čšpria pred njom, veoma je brbljiva i sklona ogovaranju.
```

⁴ The formal aspects concern the lexical, grammatical, and stylistic elements of the definitions’ language. In contrast, the substantive aspects relate to the choice of appropriate *genus proximum* and *differentia specifica*.

For better readability, we give the first part of the prompt we used in English; the example means: *imati dugačak jezik* (en: *to have a long/loose tongue*, lit. ‘to have a long tongue’) *previše pričati, biti brbljiv, ogovarati* (‘to talk too much, chatter, gossip’) *Pazi šta pričaš pred njom, ima dugačak jezik* (‘Be careful what you say in front of her, she has a big mouth’) *Pazi šta pričaš pred njom, veoma je brbljiva i sklona ogovaranju* (‘Be careful what you say in front of her, she is very talkative and prone to gossip’).

4.1.1 Gemini-2.5-Pro

The results obtained with the Gemini-2.5-Pro model are as follows:

- A1. It produced 44 idioms in the Serbian language and 6 non-idioms. Among these non-idioms, the model offered expressions void of any sense (for example, sr **baciti petao u oči*, lit. to throw a rooster in the eyes), and sometimes generated expressions that make some sense and resemble real idioms (sr **ući u vodu i ne pokvasiti se* lit. to enter the water and not get wet. If the idiom passes the first test, it can be assessed along further lines.
- A2. This model did not offer recognition of multiple senses, at least not represented in a numbered way, which may be due to our prompt, since we did not specify that multiple senses (if recognised) should be numbered. However, polysemy was identified in three cases (out of 44), but represented in one definition, separated by a comma (sr *mlatiti praznu slamu*, lit. to beat empty straw, ‘to beat around the bush’; def. ‘to engage in useless work, to talk nonsense’). Only one part of the definition was illustrated by an example: ‘They spent the whole day in the meeting beating around the bush instead of solving the problem.’ Also, an idiom sr *zabiti nož u leđa* ‘to stab someone in the back’ has the definition: ‘to betray someone, to act treacherously’ (this idiom has one more sense in sr, namely ‘to attack insidiously’, which wasn’t identified). In this case, the illustrative example covered both meanings, which may be due to their semantic connection. Sometimes the model offers false polysemy, as for the sr *terati mak na konac* (lit. to force the poppy to the thread, ‘to split hairs’) def. ‘insist on details, bring matters to their ultimate consequences’. The first part of the definition (before the comma) hits the point, but the second doesn’t.
- A3. Nine definitions that failed to describe the idiom’s meaning accurately were deemed inappropriate, and the remaining 35 were considered acceptable. For example, the idiom sr *voditi glavnu reč*, ‘to have the main say’, has a def. ‘to be the most prominent, to dominate a conversation or an activity’. It is correct in that it covers dominating in talk, but fails to explicate the fact that it is also said of a person who has the privilege to decide, to make decisions. The idiom sr *liti krokodilske suze* ‘to shed crocodile tears’ is defined using the verbs in gerund, as if it were a nominal idiom: ‘neiskreno plakanje, pretvaranje tuge’; besides, ‘pretvaranje tuge’ is an ungrammatical nominal phrase.⁵
- A4. Based on the assessment criteria, we got the following scores for the illustrative examples: out of 44 examples, 29 were assessed as good; 9 were at the very borderline

⁵ The definitions in Serbian descriptive dictionaries often contain synonyms. In these definitions, synonyms were represented relatively rarely, so we decided not to include this as a criterion for an assessment. Sometimes a synonym would enhance a definition, e.g. in the definition of an idiom sr *gledati kao tele u šarena vrata*, lit. look like a calf at a colourful door ‘to stare in amazement’, a synonym *stare* would enhance the definition. Also, in the above-mentioned idiom ‘terati mak na konac’, a synonym, *trifle*, would be a plus.

(they sounded unnatural, or contained rather unusual collocations); and six were rejected as ungrammatical and semantically inappropriate. For example, an sr idiom *španska sela* (lit. ‘Spanish villages’, ‘It’s all Greek to me’) has an appropriate definition: ‘something incomprehensible and unknow’, but is illustrated by an ungrammatical example: sr **Za mene je [to] viša matematika, to su španska sela* (lit. ‘For me [it’s] higher mathematics, it’s the Spanish villages’). sr idiom *obesiti mačku o rep* (lit. ‘to hang on a cat’s tail, ‘to brush something aside/off’) is illustrated by an ungrammatical example: sr *Sve te njegove prazne priče možeš obesiti o mačku o rep* (lit. ‘All his empty stories could be tied on a cat on a tail’), with two prepositional phrases, o+Locative, instead of one in dative and one in o+Locative.

- A5. What is interesting is the fact that the quality of the example with the idiom’s paraphrase often goes in parallel with the quality of the definition – a bad paraphrase usually indicates that the definition is not good. This is the case with the sr idiom ‘*terati mak na konac*’, mentioned above (lit. to force the poppy to the thread, ‘to split hairs’). It is paraphrased wrongly: sr ‘*Bio je toliko uporan da sazna istinu da je ispitivao svaki i najmanji trag*’. en ‘He was so determined to find out the truth that he investigated every little clue.’

However, sometimes the model generates a paraphrase that includes a detail not mentioned in the definition. As we said, the idiom sr *voditi glavnu reč*, ‘to take the lead in the discussion’, has a definition which fails to mention the fact that it is also said of a person who has the privilege to decide, to make decisions; however, this can be seen from the paraphrase: sr ‘*Na sastanku je on bio taj koji je najviše govorio i donosio odluke*’. en ‘At the meeting, he was the one who spoke the most and made decisions.’ This may indicate that the model is “aware” of this meaning but fails to explicate it in the definition.

As for the paraphrase evaluation, we decided not to assess if it sounds natural, but focused on whether it appropriately paraphrases the idiom’s meaning. Of the total of 44 paraphrases, 38 were assessed as appropriate, and six as not suitable.

4.1.2 Gemini-2.5-Flash

The results with the same prompt with this model are:

- A1. Of 50 idioms the model generated, 35 are Serbian language idioms, and 15 are false. Among the false idioms we found constructions with compositional meaning: (sr *biti u zabludi*, ‘to be mistaken’); as well as non-existent, invented idioms: (sr **biti na svojoj zemlji*, ‘to be on your own land’; sr **imati zlatan jezik*, ‘to have a golden tongue’; sr **imati kamen u stomaku*, ‘to have a stone in the stomach’, etc.).
- A2. Only in one case did the model identify a polysemy: sr idiom *baciti oko na nekoga/nešto*, lit. to throw an eye on someone/something, ‘to cast an eye on someone / at something’. This idiom was followed by a definition: ‘*Poželeći nešto, zagledati se u nekoga ili nešto*’ en ‘to desire something, to stare at someone or something’, with two existing senses (briefly, desiring and staring) presented as one, and separated only by a comma; but this way of introducing polysemy entails further problems with an example and a paraphrase, because only the first sense was illustrated/paraphrased.
- A3. Out of 35 definitions, 26 were assessed as acceptable (semantically, grammatically), and the remaining nine were assessed as unacceptable (for various reasons: either

the definition was semantically incorrect, or the wording wasn't adequate, or the definition had a grammatical error).

A4. Illustrative examples: out of 35 examples, 23 were assessed as acceptable, and 12 as not satisfactory.

A5. 25 paraphrases were assessed as good, and the remaining 10 as bad.

4.1.3 Gemini-2.5-Flash, English prompt

Notably, we also experimented with Gemini-2.5-Flash using the same prompt, but this time in English (the example we offered was in Serbian). We were wondering whether the model's reasoning would be better with an English prompt. However, the result was identical to that of the same model with the prompt in Serbian.

4.2 Experiment 2

In the second experiment, we decided to modify the prompt, assuming that we could achieve even better results, given the promising outcomes from our first experiment, particularly considering that Serbian is an under-resourced language (Krstev & Stanković, 2023). The second experiment, with the new prompt, was conducted with ChatGPT-4.1 and ChatGPT-4o.

Since we weren't satisfied with the results regarding polysemy, we emphasised that we expect the model to generate idioms with more than one sense, and also requested that separate sense definitions be listed under cardinal numbers (to avoid a single definition for multiple senses). The new prompt was (we used the prompt in Serbian, but we give it translated into English):

```
You are a lexicographer and you have to generate 50 different idioms in the Serbian language, but these idioms should have more than one meaning.
If the idiom contains a verb, and if the verb differs for verbal aspect, give an idiom with the verb in both aspects.
For each sense, you should provide a descriptive definition, and those definitions should be listed under cardinal numbers.
For each sense, give an example with the respective idiom, and another example which will illustrate the idiom's meaning, but without the idiom.
For instance: lit. to beat empty straw, `to beat around the bush'
1. to do useless work. He became a renowned painter in his later years, but when he was young, everyone believed that he was beating around the bush. He became a renowned painter in his later years, but when he was young, his painting was considered useless.
2. to talk aimlessly about the same thing, to chatter unnecessarily; to babble about something.
My grandmother has become demented and spends the whole day beating a dead horse.
My grandmother has become demented and spends the whole day rambling about nonsense.
Return the data as TSV (tab-separated values) in Cyrillic script".
```

4.2.1 ChatGPT-4.1

Here are the results obtained:

- A1. The number of idioms generated by ChatGPT-4.1 was 33 (17 rejected as non-idioms). Since many Serbian idioms with verbal elements may differ in form between the perfective and imperfective aspects, which naturally affects their definition, we believed that explicating this in the prompt would yield better results. However, it seems that this request caused unnecessary confusion, as the model generated the idiom lemma with the verb in both aspects only in six cases. In the other cases, it offered the verb in both verbal aspects as separate values. But neither were all the lemmas represented systematically in this respect, nor were all the verbs correctly represented concerning their aspect ⁶. This may explain the high number of non-idioms.
- A2. As for the request to take polysemy into account and to represent different lexical units (senses) under cardinal numbers, the results were unexpected, because the model ‘understood’ that it should generate multiple senses for each idiom and in the same way, offering idiomatic and literal meanings/definitions even if they did not exist – the model hallucinated. For instance, one of the idioms, sr *otvoriti dušu*, lit. to open the soul ‘to open up to someone’, is defined correctly in its idiomatic sense (‘to confide to someone, to tell everything to someone’), while the definition for the “second” sense was invented: ‘literally, to open an organ surgically’! Such absurd definitions were frequently offered by the model together with literal idiom definitions, which weren’t requested for (when we say that literal definitions weren’t asked for, we mean of cases of literal paraphrase of the idiom wording, and not of the idiom’s meaning: sr *biti na konju*, en ‘sitting pretty’, was defined literally: ‘to sit on horseback’). Only in one case did the model perform well, namely, offering among other idioms the one we mentioned in our prompt!
- A3. All the idioms generated by this model and assessed as genuine (33) were followed by acceptable definitions. Sometimes the definitions failed to explicate a semantic detail, but that detail would be illustrated in the example and/or paraphrase. For the idiom mentioned above, sr *biti na konju* en ‘sitting pretty’, it is specific that it stands for being in a favourable position, but that favourable position comes after some trouble; the model failed to explicate that in the definition, but represented it in example and paraphrase: *He is currently sitting pretty in his career / He is currently successful on his job* (an adverb *currently* implies that he had to overcome some difficulties).
- A4. The number of acceptable examples was 27, and of the unacceptable, 6.
- A5. The number of paraphrases assessed as good was 22 (and 11 were evaluated as bad).

4.2.2 ChatGPT-4o

The results of the second experiment with this model were:

- A1. 42 generated idioms were the Serbian language idioms (only 8 were rejected as non-existent).
- A2. This model offered 49 cases of false polysemy (except one case, the idiom with two senses from our prompt, which was repeated in the results). The senses offered as “second” were mostly cases of literal paraphrase not of the idiom meaning, but of the

⁶ In some cases the same verb was listed as only perfective (sr *dati* ‘to give’, and in some other cases as both perfective and imperfective sr *dati/davati*); in some cases false perfective/imperfective pairs or even non-existent verbs were listed (sr *mlatiti* ‘to beat/trash’/ **smlati*; sr *ići/odlaziti* ‘to go/to go away’ etc.).

idiom form, or wording; for example, sr *dati zeleno svetlo* ‘to give the green light’, is adequately defined as an idiom, ‘to permit for’, and for the “second” sense, the definition was: ‘To switch the traffic light to green’.⁷

- A3. Definitions for all idiomatic senses were assessed as acceptable, which totals 42 (plus one more, which is our definition from the prompt, repeated).
- A4. As for the illustrative examples, out of 42 examples, 29 were acceptable, and 13 were not.
- A5. Even fewer paraphrases were assessed as good, 26; the remaining 16 were rejected as bad.

4.3 Experiment 3

Since many idioms from the previous two experiments were repeated, along with their senses, definitions, and illustrative examples, we decided to conduct the third experiment, using the same prompt as in the second, but with an additional request (in a Python script) for the models to increase the temperature settings to 0.8. This experiment was done with all four models, and the results are as follows. We will present them together in Table 3 and in Figure 3.

4.3.1 Gemini-2.5-Pro, Gemini-2.5-Flash, ChatGPT-4.1, ChatGPT-4o API

The results of the third experiment were generally worse than those from the first and second experiments.

- A1. The number of non-idioms is bigger than or equal to the number of idioms, except for the Gemini-2.5-Flash, which, in general, performed best in this experiment.
- A2. It is also the only model that managed to identify two senses for five idioms. All the other models presented false polysemy.
- A3. The quality of definitions for the idioms assessed as genuine was mainly acceptable.
- A4. This also holds for illustrative examples, but to a lesser extent.
- A5. This also holds for paraphrases.

5. Results and Discussion

5.1 Experiment 1

Table 1 illustrates the results of experiment 1. The percentages in the first two rows refer to the total of 50 idioms generated; the rest of the percentages refer only to the number of existing idioms, 44 and 35, respectively.

Our first experiment yielded satisfactory results, except for the polysemy (Figure 1).

- A1. The percentage of idioms vs. non-idioms was 88%:12% for Gemini-2.5-Pro, and 70%:30% for Gemini-2.5-Flash.

⁷ In some cases, the model offered the same idiom, for example sr *držati jezik za zubima*, ‘to hold one’s tongue’, four times, with one existing, idiomatic sense, ‘to keep silent’, and three false senses (one duplicate).

		Experiment 1	
		Gemini-2.5-Pro	Gemini-2.5-Flash
This is sr idiom	yes	44 (88%)	35 (70%)
	no	6 (12%)	15 (30%),
More than one sense	yes	3 (6.81%)	1 (2.85%)
	no	41 (93.18%)	34 (97.14%)
Definition	acceptable	35 (79.54%)	26 (74.28%)
	unacceptable	9 (20.45%)	9 (25.71%)
Examples	acceptable	29 (65.9%)	23 (65.71%)
	unacceptable	15 (34.09%)	12 (34.28%)
Paraphrase	acceptable	38 (86.36%)	25 (71.42%)
	unacceptable	6 (13.63%)	10 (28.57%)

Table 1: Comparison of Gemini-2.5-Pro and Gemini-2.5-Flash in Experiment 1

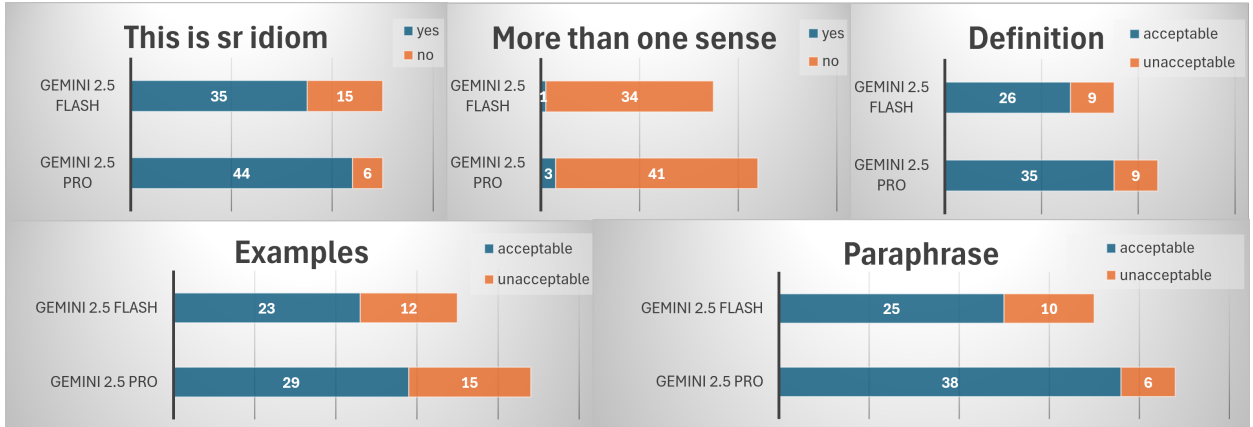


Figure 1: The experiment 1

- A2. The ratio for the polysemy vs. no polysemy is 6.81%:93.18% for Gemini-2.5-Pro, and 2.85%: 97.14% for Gemini-2.5-Flash.
- A3. The percentage of acceptable definitions was much higher than that of non-acceptable ones: 79.54%:20.45% for the Pro and 74.28%:25.71% for the Flash.
- A4. The same holds for acceptable examples vs. unacceptable; the number of acceptable examples is about two-thirds of all the examples: 65.9%:34.09% for the Pro and 65.71%:34.28% for the Flash.
- A5. Finally, the ratio of good vs. bad paraphrases was even higher in favour of good ones: 86.36%:13.63% for the Pro and 71.42%:28.57% for the Flash.

5.2 Experiment 2

Table 2 shows the results from the experiment with two ChatGPT models, with a modified prompt. What we said in 5.1 for the percentages in the first two rows and in the rest also holds here.

		Experiment 2	
		ChatGPT-4.1	ChatGPT-4o
This is sr idiom	yes	33 (66%)	42 (84%)
	no	17 (34%)	8 (16%)
More than one sense	yes	false polysemy	
	no		
Definition	acceptable	33 (100%)	42 (100%)
	unacceptable		
Examples	acceptable	27 (81.81%)	29 (69.04%)
	unacceptable	6 (18.18%)	13 (30.95%)
Paraphrase	acceptable	22 (66.66%)	26 (61.9%)
	unacceptable	11 (33.33%)	16 (38.09%)

Table 2: Results from Experiment 2 for ChatGPT-4.1 and ChatGPT-4o

The second experiment was conducted with ChatGPT-4.1 and 4o, with a prompt similar to the one in the first experiment, but with an additional request to consider the verbal aspect, as well as a specific request to generate idioms with multiple senses. Even though we provided an example idiom to guide the models towards the desired result, it seems that our prompt confused the models.

The overall results are slightly worse than with the first prompt.

- A1. The percentage of idioms vs. non-idioms was 66% vs. 34% for ChatGPT-4.1, and 84%:16% for ChatGPT-4o.
- A2. Both models struggled with false polysemy, without offering even one case of existing multiple senses.
- A3. All the definitions were assessed as acceptable, a rather interesting result.
- A4. High percentages of examples are assessed as good for both models: 81.81% for ChatGPT-4.1 and 69.04% for ChatGPT-4o.
- A5. The same holds for paraphrases; although, these percentages for good paraphrases are lower: 66.66% and 61.9% for ChatGPT-4.1 and 4o respectively.

The results from Table 2 can be most effectively visualized as grouped bar charts (Figure 2), where each model is represented along the x-axis and the y-axis shows the number of true idioms (yes) and false generation (no), acceptable and unacceptable outputs for definitions, examples, and paraphrases.

5.3 Experiment 3

Table 3 shows the results from the third experiment, conducted via Python scripts, with an additional request for the models to increase the temperature settings to 0.8. Percentages in the first two rows are per the total number of idioms produced by models (50 in all cases, except in the case of ChatGPT-4.1, which produced 47). The remaining percentages refer only to the existing idioms.

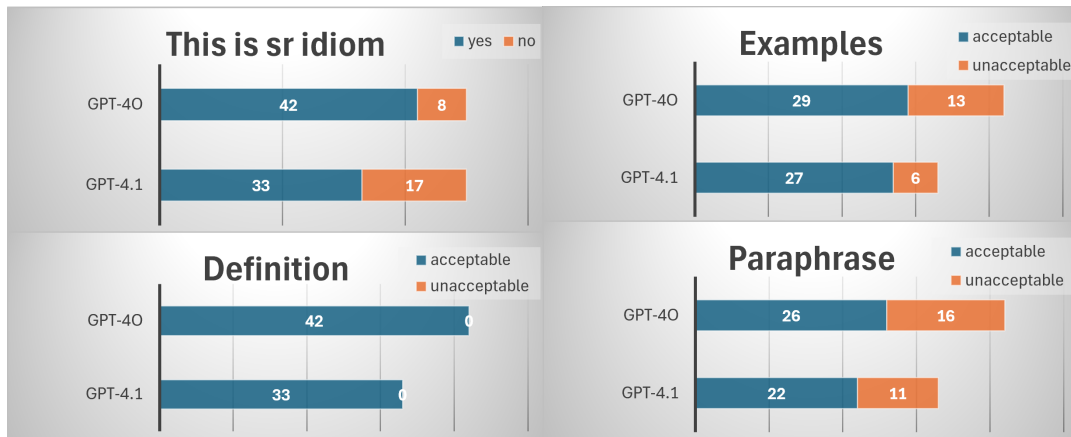


Figure 2: The experiment 2

- A1. As the table shows, the third experiment resulted in higher percentages of non-idioms. In this respect, the results of the third experiment are the worst.
- A2. Except Gemini-2.5-Flash, which offered five idioms with two senses, all the models struggled with false polysemy.
- A3. The percentage of acceptable definitions was rather high for all the models.
- A4. The quality of the examples, as indicated by the percentages, wasn't satisfactory.
- A5. It seems that the quality of paraphrases follows the low quality of examples.

The results from Table 3 can be most effectively visualised as grouped bar charts (Figure 3), where each model is represented along the x-axis. The y-axis shows the number of true idioms (yes) and false generation (no), as well as acceptable and unacceptable outputs for definitions, examples, and paraphrases.

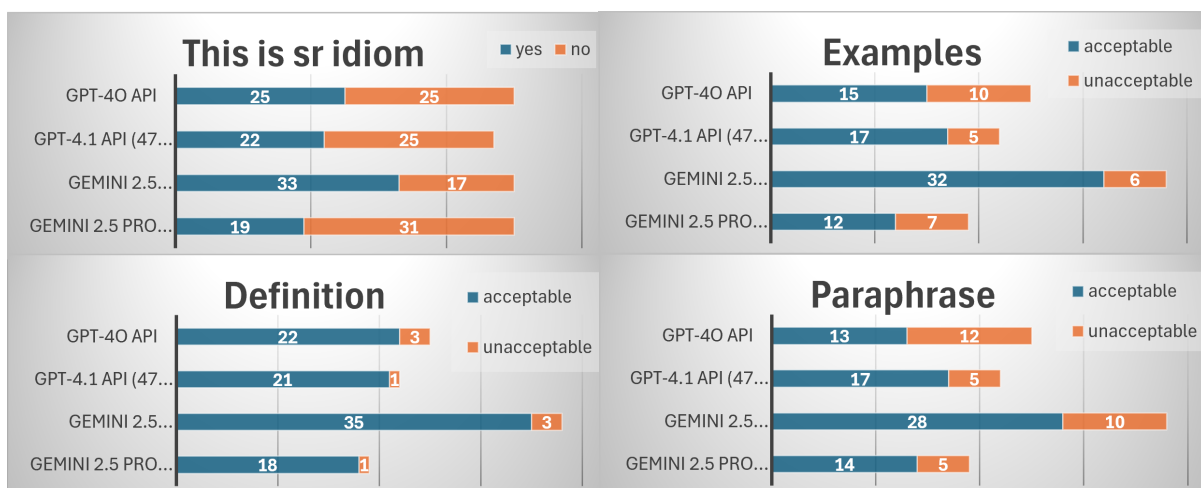


Figure 3: The experiment 3

6. Conclusion

This study provided a comprehensive evaluation of the capabilities and limitations of state-of-the-art large language models (LLMs) for idiom identification, sense disambiguation,

	Experiment 3			
	Gemini-2.5 -Pro API	Gemini-2.5 -Flash API	ChatGPT-4.1 API (47 idioms)	ChatGPT-4o API
This is sr idiom				
yes	19 (38%)	33 (66%)	22 (46.8%)	25 (50%)
no	31 (62%)	17 (34%)	25 (53.19%)	25 (50%)
More than one sense				
yes	false polysemy	5 (15.15%)	false polysemy	false polysemy
no		28 (84.84%)		
Definition		38 lexical units		
acceptable	18 (94.73%)	35 (92.1%)	21 (95.45%)	22 (88%)
unacceptable	1 (5.26%)	3 (7.89%)	1 (4.54%)	3 (12%)
Examples				
acceptable	12 (63.15%)	32 (84.21%)	17 (77.25%)	15 (60%)
unacceptable	7 (36.84%)	6 (15.79%)	5 (22.72%)	10 (40%)
Paraphrase				
acceptable	14 (73.68%)	28 (73.68%)	17 (77.27%)	13 (52%)
unacceptable	5 (26.31%)	10 (26.32%)	5 (22.72%)	12 (48%)

Table 3: Results for Experiment 3

definition generation, and the creation of illustrative examples in Serbian. To the best of our knowledge, this is the first piece of research on this topic for the Serbian language, which builds on (Marković & Stanković, 2025). Our experiments with Gemini-2.5-Pro, Gemini-2.5-Flash, ChatGPT-4.1, and ChatGPT-4o revealed that, while LLMs can reliably generate a considerable number of genuine Serbian idioms and acceptable definitions, significant challenges remain.

Firstly, all tested models exhibited a notable tendency to produce non-idioms or compositional phrases, with the rate of false positives increasing when prompts were made more complex or temperature settings were raised. Polysemy was poorly handled: models frequently generated false or literal senses instead of accurately capturing true idiomatic polysemy, suggesting a lack of deep semantic understanding for multiword expressions. In this respect, our findings are consistent with those from recent studies for English, namely that LLMs struggle with handling even simple polysemic structures.

Definition quality was generally high for idioms correctly identified by the models. Some definitions were erroneous (with grammatical errors, modelled for inappropriate part of speech, etc.). However, our findings in this respect confirm what has been observed in previous studies for English, namely that it is evident that models generate definitions using formulae derived from dictionaries used for their training. The models’ performance in generating illustrative examples was more variable — while the majority were rated as natural and appropriate, a non-trivial share suffered from unnatural collocations or ungrammatical constructions. Paraphrase generation also proved to be a challenging aspect,

with a substantial number of paraphrases failing to preserve the original idiomatic meaning or providing awkward formulations.

Our results confirm earlier findings in the literature that LLMs are prone to literal interpretations or hallucinated senses for idiomatic language. Even with carefully engineered prompts, current LLMs have only a partial grasp of idiomaticity, and their "understanding" of idioms appears largely pattern-based rather than rooted in genuine semantic modelling. This is particularly evident in under-resourced languages, such as Serbian, where training data may be limited.

Nevertheless, the models showed promising results in assisting lexicographic workflows — especially in drafting definitions and generating examples — which can be valuable for accelerating dictionary development when combined with human expertise and careful post-editing. Our findings underscore the importance of leveraging both AI and expert linguistic knowledge, suggesting that further progress will require larger and more diverse training corpora for Serbian, as well as the continued refinement of prompt design and evaluation protocols.

In summary, while LLMs are already "so close" to providing meaningful support for phraseological lexicography in Serbian, significant gaps remain that must be addressed. Future research should focus on improving sense disambiguation, reducing hallucinations, and developing targeted evaluation benchmarks for idiomatic expressions in low-resourced languages. Additionally, various aspects of idiomatic MWEs should be considered, including their frequency in Serbian language corpora, register, domain, and so on.

Additionally, one should consider processing speed and usage costs: older models are often more affordable than newer ones, and if they adequately fulfil the task, they might better align with the project's budget.

7. Acknowledgements

This work was supported by the Ministry of Science, Republic of Serbia #GRANT 451-03-136/2025-03/200174, and the Science Fund of the Republic of Serbia (#7276, Text Embeddings – Serbian Language Applications, TESLA); COST ACTION CA21167 - Universality, Diversity, and Idiosyncrasy in Language Technology (UniDive).

8. References

- Beliga, S. & Filipović Petrović, I. (2024). Large Language Models Supporting Lexicography: Conceptual Organization of Croatian Idioms. In *Proceedings of the Conference on Language Technologies and Digital Humanities*. pp. 23–46.
- de Schryver, G.M. (2024). The Road towards Fine-Tuned LLMs for Lexicography. *Large Language Models and Lexicography*, p. 6.
- de Schryver, G.M. & Joffe, D. (2023). The end of lexicography, welcome to the machine : on how ChatGPT can already take over all of the dictionary maker's tasks. URL <https://www.youtube.com/watch?v=mEorw0yefAs>.
- Filipović-Petrović, I. & Beliga, S. (2024). Lexicographic Treatment of Idioms and Large Language Models: What Will Rise to the Surface? In *Book of Abstracts of the Workshop Large Language Models and Lexicography*. Ljubljana: Centre for language resources and technologies, University of ..., pp. 12–16.

- Gantar, A. (2024). Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica. In S. Marjanović (ed.) *Moderni rečnici u funkciji prosečnoga korisnika: stari problemi, savremeni pravci i novi izazovi*. University of Belgrade, Faculty of Philology, pp. 151–157.
- Gortan Premk, D. et al. (1959–2023). *Rečnik srpskohrvatskog književnog i narodnog jezika SANU, I–XXII [The Dictionary of the Serbo-Croatian Standard and Vernacular Language]*. Beograd: Institut za srpski jezik SANU i SANU. 22 volumes.
- Jakubiček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. pp. 518–533.
- Kilgarrieff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, volume 1. Universitat Pompeu Fabra Barcelona, pp. 425–432.
- Kovačević, Ž. (2002). *Srpsko-engleski frazeološki rečnik*. ” Filip Višnjić”.
- Krstev, C. & Stanković, R. (2023). *European Language Equality: A Strategic Agenda for Digital Language Equality*, chapter Language Report Serbian. Cham: Springer International Publishing, pp. 203–206. URL https://doi.org/10.1007/978-3-031-28819-7_32.
- Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1), pp. 1–10.
- Marković, A. & Stanković, R. (2025). Primena velikih jezičkih modela u srpskoj opisnoj leksikografiji – studija slučaja. In M. Dinić Marinković & B. Kovačević (eds.) *Applied Linguistics Today – Modern Approaches to Old and New Challenges*. University of Belgrade – Faculty of Philology. Accepted.
- Otašević, Đ. (2012). *Frazeološki rečnik srpskog jezika*. Prometej.
- Phoodai, C., Rikk, R., Medved, M., Měchura, M., Kosem, I., Kallas, J., Tiberius, C. & Jakubiček, M. (2023). Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner’s Dictionary within the Microstructural Framework. In *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex2023 conference*. pp. 345–375.
- Rundell, M. (2024). Automating the creation of dictionaries: are we nearly there? *Humanising Language Teaching*, 26(1).
- Stanković, R., Šandrih, B., Stijović, R., Krstev, C., Vitas, D. & Marković, A. (2019). SASA dictionary as the gold standard for good dictionary examples for Serbian. *Electronic lexicography in the 21st century: Smart lexicography*, pp. 248–269.
- Stevanović, M. et al. (ed.) (1982). Matica srpska; Matica hrvatska, fototipsko izd. edition. Knjige 1, 2, 3 - zajedničko izd. Matice srpske i Matice hrvatske; knj. 4, 5, 6 - izdanje Matice srpske.

This work is licensed under the Creative Commons Attribution ShareAlike 4.0 International License.

<http://creativecommons.org/licenses/by-sa/4.0/>

