LARGE LANGUAGE MODELS SUPPORTING LEXICOGRAPHY: CONCEPTUAL ORGANIZATION OF CROATIAN IDIOMS

Slobodan BELIGA,[1,2] Ivana FILIPOVIĆ PETROVIĆ[3]

[1]University of Rijeka, Faculty of Informatics and Digital Technologies, Rijeka, Croatia

[2]University of Rijeka, Center for Artificial Intelligence and Cybersecurity, Rijeka, Croatia

[3]Croatian Academy of Sciences and Arts, Zagreb, Croatia

In this paper, we describe how large language models respond to queries on the semantic features of idiomatic expressions in Croatian. Specifically, we created queries for four large language models using a sample of 430 idioms from the freely available *Online Dictionary of Croatian Idioms*. These idioms were previously categorized into 65 concepts or semantic categories. Since this work was done manually by linguists and lexicographers, we wanted to investigate the quality and accuracy of the results obtained by artificial intelligence-based systems and compare them with those obtained by human intelligence. The aim was to assess whether the responses are suitable for lexicographic processing and to what extent lexicographers can use them, possibly as a reliable tool for the automatic creation of a conceptual organization of idioms.

Keywords: large language models, idioms, semantic similarity, conceptual organization in lexicography

1 INTRODUCTION

The rapid advancement of artificial intelligence (AI) impacts nearly all areas of knowledge and society, and lexicography is no exception. The reflection of social and technical revolutions in dictionaries is not a new phenomenon in this discipline. Technological developments in the form of various tools, corpora, dictionary writing systems, and user interfaces have been eagerly anticipated and embraced by lexicographers. The emergence of AI, particularly large language models (LLMs), has raised numerous questions about its impact on lexicography (cf. de Schryver, 2023). These questions range from whether previous technologies and lexicographical methods can be abandoned to investigating which lexicographical tasks might benefit from AI. Lew (2023) and Tran et al. (2023) explore how to integrate linguistic and lexicographic human knowledge with the latest advances in AI technology, specifically LLMs to determine their current utility for lexicographic purposes.

Before the rise of advanced AI tools like ChatGPT,[1] semi-automatic dictionary creation based on the model of post-editing lexicography gradually became a desired standard, inspired by successful initial projects (Baisa et al., 2019; Jakubiček et al., 2021; Kosem et al., 2014). In particular, the creation of contemporary e-dictionary relies on two foundations: technology that automatically performs many steps in dictionary-making, and the post-editing work of lexicographers and linguists who manually evaluate the results, refine parts of the entry, and finalize its a ppearance. With the emergence of AI tools, there is a growing advocacy for human lexicographers to collaborate with generative AI chatbots like ChatGPT in creating dictionaries. According to some opinions, this collaboration may render concordances, keywords, and other corpus-based technologies obsolete (Fuertes-Olivera, 2024). The approach is promoted as being more efficient, cost-effective, and capable of retrieving hard-to-obtain data. Since conclusions about speed and costs can only be made after the completion of the dictionary-making project, in this paper, we will focus on the idea that AI tools can perform specific tasks that would otherwise require significant human resources and will lead to data that are harder to obtain.

Opinions that express a negative attitude towards the quality of AI technologies and criticize their use (see Vossen, 2022) raise concerns about the potential for widespread hallucinations. They also highlight concerns regarding the accuracy of data and the level of trust that users place in the data provided by AI. In that context, Rundell (2023) emphasizes the importance of dictionaries because they are associated with confidence that the information in the dictionary is accurate, which has been drawn from the Enlightenment when they taught about proper usage, and the idea of ’accurate’ data stored in dictionaries has become established in the minds of users (Filipović Petrović, 2018). If we rely on Hargraves’ thought, presented in (Hargraves, 2018), that we are facing a gap between the great availability of big data and the addressability of that data, we can conclude that post-editing lexicography remains an indis-

[1] ChatGPT is a chatbot and virtual assistant developed by OpenAI (launched on November 30, 2022).

pensable condition for creating dictionaries, no matter which technology we use to obtain linguistic data.

In this paper, we use LLMs and ChatGPT to automate the process of finding semantic equivalents (task one) among the idioms contained in the *Online Dictionary of Croatian Idioms*[2] (ODCI). We also use them to automatically generate semantic fields or concepts to which the idioms belong (task wo). Since linguists have done this work manually on a certain number of idioms, we test large language models and ChatGPT to examine the quality and accuracy of their results compared to those resulting from human intelligence. However, it should be noted that in this instance, human intelligence was used to organize concepts from a lexicographical standpoint, considering the dictionary type and the potential needs of users. The research aims to find a reliable tool for the automatic creation of a conceptual organization of idioms based on data from ODCI.

We believe it is worth testing the capabilities of LLMs in this research for several reasons. First, we have a crucial starting point: human input from lexicographers' knowledge and introspection, as well as human evaluation of equal quality. Additionally, the research we have chosen faces the challenge of identifying data that is difficult for existing technologies to p rocess. For example, Google Translate service often translates Croatian idioms word-for-word, without considering their actual meanings. Previous work by Moussallem et al. (2018) and Filipović Petrović et al. (2024) has made valuable contributions to linking idioms from different languages based on semantic similarity, but these studies are based on small datasets. Finally, achieving the desired result of conceptually organizing Croatian idioms is unlikely without the use of automation, and expecting it to be accomplished quickly enough to justify the effort is not realistic.

This paper is structured as follows. After the introduction, in Section 2 we describe the linguistic resource on which we base our research and give a theoretical overview of conceptual organization in lexicography. Then, in Chapter 3, we have described the experiments we conducted with LLMs and the results we obtained. Finally, the conclusion follows.

[2] https://lexonomy.elex.is/#/frazeoloskirjecnikhr

2 THE ONLINE DICTIONARY OF CROATIAN IDIOMS: TECHNOLOGY AND POST-EDITING LEXICOGRAPHY

Although developments in the application of language technologies to the compilation of dictionaries looked quite promising in the past ten years, a considerable number of European languages remain low or middle-resourced (Rehm & Way, 2023). This is also true for Croatian, especially in terms of freely available e-dictionaries and resources. The project[3] of creating the Online Dictionary of Croatian Idioms was launched in 2019 at the Croatian Academy of Sciences, relying on freely accessible lexicographic tools and lexicographers with a linguistic background. The goal was to develop an open-access born-digital dictionary based on a corpus, and we made efforts to implement a post-editing lexicography model. In this model, the role of lexicographers is to evaluate and refine, i.e., post-edit data that has been generated automatically and transferred into a dictionary writing system. This is not entirely the case with this Croatian dictionary, but several separate automated processes have been utilized. For corpus searches, we used Sketch Engine, which was freely available to members of the academic community within the ELEXIS project from 2018 to 2022. It served as a tool for obtaining concordances from hrWaC, the largest Croatian corpus at the time of its release (Ljubešić & Klubička, 2016). Furthermore, Lexonomy,[4] a platform designed for creating and publishing dictionaries, served as both a dictionary writing system and a user interface. The lexicographic processing involved a combination of manual and automated methods. Concordances were manually scanned and analyzed, and tools like Word Sketch were used to extract multiword expressions. Additionally, frequency and typical usage statistics were employed, specifically using the LogDice metric, which measures the strength of association between words in collocations, helping to identify commonly co-occurring terms. The GDEX (Good Dictionary Example) algorithm was also utilized to select the most representative and illustrative usage examples from the corpus, ensuring that the examples provided in the dictionary are both typical and informative. Entries in Lexonomy were added manually. Version 2 was released in 2023, contains 563 entries and 1165 idioms (Filipović Petrović & Parizoska, 2023).

[3] https://frazeoloski-rjecnik.eu/en/
[4] https://lexonomy.elex.is/

Given that lexicography is a constant race against time, the next aim of the project was to produce more content within a shorter timeframe. Thus, we turn to the automatic identification of idioms in the newly created larger and more contemporary corpus, the CLASSLA corpus for Croatian (Ljubešić et al., 2024; Ljubešić & Kuzman, 2024). The intention was to gather comprehensive lists of automatically recognized idioms (Filipović Petrović & Kocijan, 2024; Kocijan et al., 2023) and then perform post-editing lexicography, as well as produce standalone resources such as datasets that can be reused in NLP, machine translation, and cross-lingual studies.

2.1 Conceptual Organization

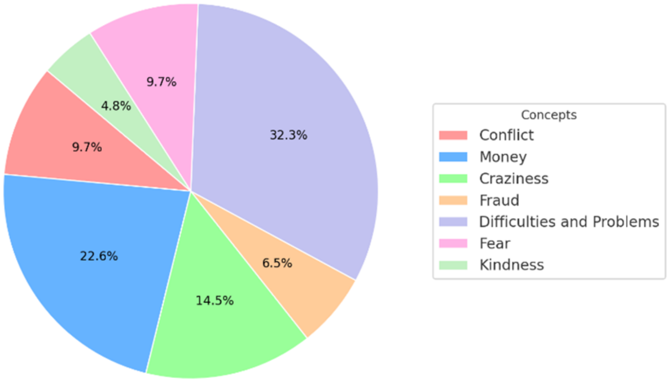
When it comes to the presentation of dictionary content in digital media, technology has also opened up many possibilities. The introduction of hyperlinks connecting alphabetically distant entries, such as multi-word expressions that have similar meanings but different structures and forms, would probably deeply impress lexicographers who, before the computer age, made endless lists of domains of human knowledge, trying to categorize expressions into ideas and make them searchable in the linear format of printed media. For phraseological dictionaries in particular, the ability to link idioms that are very different in expression can be considered quite revolutionary. For example, idioms such as *fali komu daska u glavi* (lit. someone is missing a plank in the head) and *nisu komu sve koze na broju* (lit. not all goats are on someone's count) both mean 'to be crazy or insane.' When the user looks up these idioms in a printed dictionary, he may find them only under the first noun in the construction, such as plank or goat, without knowing that the other idiom exists. For this reason, lexicographers have always looked for ways to represent semantically related words, even though alphabetical order dominates in most dictionaries. Proponents of this approach believed that conceptual organization better fits the way the human mind organizes its ideas and words, and that lexicography should not only help users find the meaning starting from the words but also the words and expressions starting from the idea or concept (Geeraerts, 1989; McArthur, 1986).

With this in mind, the conceptual organization was manually compiled based on the content of the ODCI. Sixty-four concepts were designed, into which

430 idioms have been categorized so far. In manually creating concepts and grouping idioms into them, the lexicographers working on this phraseologi-cal dictionary primarily relied on examples of good lexicographical practice, as found in dictionaries such as the Collins COBUILD Idioms Dictionary (2002) and the Cambridge Idioms Dictionary (2006). These dictionaries organize idioms alphabetically but also have special sections where idioms are grouped by common themes such as love, honesty, deception, disagreement, success and failure, progress, happiness, and sadness. This idea of conceptual organization in lexicography, which consists of dividing knowledge about the world into thematic areas such as life, body, people and community, construction, emotions and attitudes, thought and communication, materials, objects, equipment, art, science, technology, industry, education, entertainment, transportation, and abstract concepts, is found in the most famous thesauruses, such as Roget's Thesaurus of English Words and Phrases (1852), and has been adapted over time to the nature of the dictionary and the spatial lim-itations of the medium in which it is published. A word or idiom can be catego-rized into several concepts, and human intelligence, beliefs, knowledge, and instincts guide the decision to do so. This is important as in this paper we focus on what artificial intelligence will produce compared to human intelligence. The criterion for connecting semantically similar idioms in ODCI is to find com-mon semantic and structural elements within the idioms, which are further ex-plained in Filipović Petrović and Parizoska (2019). As a sample of the manually crafted conceptual organization for the ODCI, we selected 7 concepts, and the diagram in Figure 1 illustrates the frequency distribution of idioms across these concepts. Notably, concepts such as difficulties/problems, money and conflict stand out as semantic fields rich in expressive idiomatic expressions. The entry in the ODCI includes links to semantically related idioms. Additionally, a separate resource has been created: a conceptual index containing a list of concepts and corresponding idioms that serve as links to entries in the main dictionary. This enables users to search by concepts, starting from the idea. Creating the conceptual index was an extensive and demanding task in terms of human resources and time and it is also prone to oversights.

The ODCI will continue to be augmented with new entries as corpus research and automatic identification progress. This is why further technological im-

Figure 1: Example of the distribution of seven concepts in the Online Dictionary of Croatian Idioms.



provements are being pursued to automate conceptual organization. The pro-cess should be conducted at three levels.

1. On the existing material. The goal is to find concepts for the remaining uncategorized idioms. This involves determining whether they fit into existing concepts based on their meaning but were overlooked during manual organization, or proposing new concepts that they belong to. Each idiom should be assigned to a specific concept based on its meaning, even if it initially stands alone within that concept. This approach is valuable for future additions to the dictionary, as it allows for new idioms to be grouped under the same concept. As a result, users will be able to search the dictionary by ideas and meanings, with some concepts containing multiple idioms and others just one. Over time, as new idioms are added, these concepts may evolve and expand.

2. On the new material. As mentioned, the dictionary will be supplemented with new entries, which also means new meanings. Based on this, concepts corresponding to those meanings can be found and additional idioms will be associated with them.

3. For new idioms that do not fit into existing concepts, new concepts would be proposed, expanding the list of entries in the conceptual index.

To achieve this, in this research, we conducted a pilot study on several LLMs and several manually crafted concepts from the ODCI. First, we conducted two test experiments, and then we assigned two tasks to the AI system that proved to be successful in the tests. In the next section, we describe the procedures we conducted.

3 LEXICOGRAPHY AND LARGE LANGUAGE MODELS: LET'S GIVE IT A TRY

We wanted to examine how existing LLMs, open-source and available for use, can help determine semantic similarity and automatically build idiom lexicons with their associated semantic fields. In the experiments, we used 3 different LLMs, namely Cro-CoV-cseBERT, bcms-bertic, and gpt2-vrabac.

We used the Cro-CoV-cseBERT[5] model (Babić et al., 2021) which is based on the CroSloEngualBERT model (cseBERT) (Ulčar & Robnik-Šikonja, 2020), and fine-tuned on a large corpus of texts related to the COVID-19 in the Croat-ian language. It is important to emphasize that CroSloEngualBERT is a trilin-gual BERT-based language model that was pre-trained on a large volume of texts from online news articles in Croatian, Slovene and English (5.9 billion to-kens; comprising 31% Croatian, 23% Slovenian, and the remaining portion in English), and is additionally fine-tuned only with Croatian corpora from a specific domain covering the COVID-19 topic (186,738 news articles and 500,504 user comments related to COVID-19 published on Croatian online news por-tals and 28,208 COVID-19 tweets in the Croatian language). Other fine-tuning details are described in (Babić et al., 2021). Cro-CoV-cseBERT is fine-tuned for the masked language modeling task.

The next model we used is bcms-bertic[6] (BERTić). It is a transformer model pre-trained on 8 billion tokens of crawled text from Croatian, Bosnian, Serbian and Montenegrin web domains (Ljubešić & Lauc, 2021). bcms-bertic was

[5] https://huggingface.co/InfoCoV/Cro-CoV-cseBERT
[6] https://huggingface.co/classla/xlm-r-bertic

trained using the ELECTRA transformer approach. In terms of dimensions, the BERTić and the cseBERT belong to the base-sized models.[7]

In addition to the mentioned BERT and ELECTRA architectures, we also wanted to test the efficiency of a model that uses a third type of architecture, namely the Generative Pre-trained Transformer (GPT) and its size is also smaller. Con-sidering the kinship of the Croatian and Serbian languages, we also tested an available generative model for the Serbian language called gpt2-vrabac[8] (Škorić, 2024). The model was trained on about 4 billion tokens. The model gpt2-vrabac has 136 million parameters and is based on the GPT2-small ar-chitecture. The model was trained on datasets containing the texts of doctoral dissertations, a corpus of public discourse in the Serbian language, corpora containing texts from the web, and the corpus of the Society for Language Re-sources and Technologies. More details can be found in (Škorić, 2024).

The selected LLMs are employed to compute the semantic similarity between a specified semantic field, i.e. concept (e.g., kindness), and the entire corpus of Croatian idiomatic expressions available in the lexicon. For each idiomatic expression drawn from the lexicon and the specified semantic field, a tokenization process is applied to segment the lexical units into discrete tokens, subsequently forwarded to the language model for embedding extraction. Following this, the resultant vectors of all tokens are aggregated to yield a singular vector, subsequently normalized by the token count within the sequence. This methodology thus furnishes the averaged vectorial representation of the lexical item or idiomatic expression (i.e., the centroid-averaged token vectors approach). In this procedural framework, both the semantic field and the entire spectrum of idiomatic expressions sourced from the lexicon undergo computation for the collective embedding of all constituent tokens, thus eliciting a unique vector for each idiomatic expression and a distinct vector for the semantic field. After the derivation of embeddings for the semantic field and the corpus of idiomatic expressions, cosine similarity is deployed to quantify the degree of semantic correspondence between the semantic field and each

[7] Base-sized language models typically have 12 hidden layers and around 110 million parameters (Ljubešić & Lauc, 2021), such as BERT-base and XLM-R base. In contrast, large-sized language models often feature 24 or more hidden layers and can range from hundreds of millions to billions of parameters (Clark et al., 2020), examples being BERT-large, GPT-3, and XLM-R large.

[8] https://huggingface.co/jerteh/gpt2-vrabac

idiomatic expression. Higher cosine similarity scores denote heightened congruity between the vectors (i.e., the semantic field and the idiomatic expression).

In addition to leveraging freely available and open-source models, our investigation extended to evaluating the efficacy of a substantially larger and commercially developed model, namely the GPT-3.5-turbo model, for the task of matching all idioms within the same semantic field. Employing prompt engineering methodologies, we conducted a thorough examination of the performance of both the ChatGPT and the GPT-3-turbo model developed by OpenAI (Brown et al., 2020). The GPT-3.5-turbo model utilizes a transformer-based architecture with 175 billion parameters. The model has 96 transformer layers, each vector representation within the model has 12,288 dimensions (hidden states), and there are 96 attention heads in each layer. With such a specification, the model is extremely powerful in recognizing and generating complex patterns in the text. Thus, GPT-3.5-turbo significantly surpasses the base-sized language models in terms of size and model capacity (for the comparison: BERTić and cseBERT have 12 hidden layers and 768 hidden states). However, although not trained on corpora of Croatian texts, Perak et al. (2024) showed that the OpenAI GPT model for the Croatian language provides satisfactory results with prompt engineering techniques for the causal commonsense reasoning task for the Croatian language, even when it came to dialectal (DIALECT-COPA)[9] rather than the standard Croatian language.

It is important to note several experimental specifications related to the use of commercial models. The experiment was conducted in March and April 2024. Although the GPT-4 model had an available Application Programming Interface (API) at that time, we utilized the more cost-effective GPT-3.5 turbo model. The GPT-3.5 turbo is approximately ten times cheaper than the GPT-4 for both input and output tokens. Tokens can be thought of as pieces of words, where 1,000 tokens[10] correspond to about 750 words. The context window for the GPT-3.5 turbo model is 16,385 tokens, whereas the basic GPT-4 version has a

[9] In COPA task (Choice of Plausible Alternatives) a model has to select which of the two candidate statements are more likely to be the cause or effect of a given premise statement.

[10] According to OpenAI specifications available at official OpenAI website (August 2024): https://openai.com/api/pricing/.

context window of 8,192 tokens. Given these considerations, particularly the cost, we selected the GPT-3.5 turbo model for this initial research.

Considering the rapid development dynamics of the GPT model, currently (August 2028) newer GPT-4o and GPT-4o-mini models are available. These models feature a context window size of 128K tokens, execute quickly, and exhibit higher intelligence than GPT-3.5 turbo. The cost[11] of the GPT-4o model is $5 per 1M input tokens and $15 per 1M output tokens, while the GPT-4o-mini costs $0.15 per 1M input tokens and $0.6 per 1M output tokens. Just four months later, we could repeat the same experiment for the same cost, but using a model that is significantly faster, larger in terms of parameters and the corpus on which it was trained, and better suited for the Croatian language. This is because it is adapted to multiple languages beyond English, utilizes a significantly larger context window, and is trained on data up to October 2023.

3.1 Experiment one

From the manually created conceptual organization in the ODCI, the following samples were selected for the research: a list of 150 idioms distributed across 27 concepts. For testing LLMs, three concepts from the conceptual index list were selected: kindness, madness and conflict. Table 1 shows the selected concepts and their corresponding idioms.

In the first experiment, we used LLMs to calculate the semantic similarity of idioms and the given semantic field. The task was set so that from a list of 150 idioms algorithm finds those that by meaning belong to the following semantic fields or concepts: 1) kindness, 2) madness and 3) conflict. LLMs like Cro-CoV-cseBERT, bcms-bertic, and gpt2-vrabac have, on average, ranked three idioms belonging to the concept of kindness between 47th and 65th place. The best result was achieved by gpt2-vrabac for the idiom *duša od čovjeka* (lit. soul of a person) ‘a kind person’, placing it in 5th place. They ranked the idioms *zlatna koka* (lit. golden goose) ‘cash cow’, *mala beba* (lit. little baby) ‘something easy to use, harmless’ and *malo sutra* ‘no way, no chance’ in the first place.

[11] All listed prices were taken from the official OpenAI site (https://openai.com/api/pricing/) on August 15, 2024.

Furthermore, for the concept madness, the Cro-CoV-cseBERT model placed *zreo za ludnicu* (lit. ripe for the madhouse) in the first place, *lud kao šiba* ‘crazy like a hatter’ in 5th place, and *lud sto gradi* ‘crazy like a hundred’ in 10th place. The gpt2-vrabac placed *lud sto gradi* in 5th place, *zreo za ludnicu* in 6th place, and *lud kao šiba* in 13th place, while bcms-bertic placed *lud sto gradi* in 22nd place, with all other idioms being further ranked.

Table 1: Concepts and corresponding idioms involved in Experiment one.

Concept	*Idioms*
`kindness`	*dobar kao kruh* (lit. as good as bread) ‘very good, hearted’, *duša od čovjeka* (lit. soul of a person) ‘a kind person’, *ne bi ni mrava zgazio* ‘wouldn’t hurt a fly’
`madness`	*fali daska u glavi komu* (lit. someone is missing a plank in the head) ‘not normal’, *lud kao šiba* ‘crazy like a hatter’, *lud sto gradi* ‘crazy like a hundred’, *nisu sve koze na broju komu* (lit. not all the goats are in the pen) ‘crazy, not normal’, *nisu svi doma komu* (lit. not everyone is at home) ‘crazy, not normal’, *posvađao se s mozgom* (lit. quarreled with the brain) ‘lost one’s mind’, *zreo za ludnicu* (lit. ripe for the madhouse), *puknuti kao kokica* (lit. to pop like a popcorn) ‘go crazy’, *najesti se ludih gljiva* (lit. to eat mad mushrooms) ‘go crazy’
`conflict`	*dolijevati ulje na vatru* (lit. to pour oil on the fire) ‘further inflame a conflict or disagreement’, *izvrijeđati na pasja kola koga* ‘to verbally abuse someone thoroughly’, *lome se koplja* (lit. spears are breaking) ‘there’s a fierce conflict’, *posijati sjeme razdora* (lit. to sow the seeds of discord), *posvađati se na mrtvo ime* ‘to fight bitterly’, *posvađati se na pasja kola* ‘to fight fiercely’, *stvarati zlu krv* (lit. to create bad blood), *svađati se kao pas i mačka* (lit. to fight like cats and dogs), *prosipati žuč* (lit. to spill bile) ‘to express bitterness’, *spaliti mostove* (lit. to burn bridges), *ukrstiti koplja* (lit. to cross swords) ‘to engage in a conflict’

Finally, for the field `conflict`, model bcms-bertic placed the idiom *stvarati zlu krv* (lit. to create bad blood) in 8th place, gpt2-vrabac placed the idiom *lome se koplja* (lit. spears are breaking) ‘there’s a fierce conflict’ in first place and *ukrstiti koplja* (lit. to cross swords) ‘to engage in a conflict’ in 6th place, while Cro-CoV-cseBERT ranked the idiom *prosipati žuč* (lit. to spill bile) ‘to express bitterness’ highest, placing it in 24th place. In this ranking, a lower number indicates a better result. For example, if an idiom is ranked in the first place, it means the system considers it the best match for the given concept of `kindness`. Conversely, rankings of 47th and 65th suggest that the system considers those idioms to be a poor match for the concept. Despite several successfully ranked idioms paired with predefined concepts, the overall results for all idioms in Table 1 are not good enough to be useful in lexicographic work.

The examined LLMs for Croatian do not yield high-quality results for figurative language. They use different types of texts in the model training process. For example, BERTić is trained on a large text corpus that includes various types of content, including web pages, literary works, and newspaper articles (Ljubešić & Lauc, 2021). Although the training corpus is not specifically designed for idioms, it naturally includes many idiomatic expressions that appear in every-day language. However, it seems that this quantity of idioms is not quite suf-ficient for LLM to be efficient for our lexicographic task. This suggests that there is significant room for improvement in this area. For instance, selecting a corpus richer in idiomatic expressions when creating a model for idioms, and employing techniques such as fine-tuning, transfer learning, or other methods for model enhancement, could be beneficial. In addition, Croatian is currently under-resourced in terms of a large corpora rich in idiomatic expressions, and which is vital for training language models to enhance their performance for our lexicographic task. Besides, the issue of multi-word constructions not repre-senting the sum of the meanings of their parts is a well-known challenge in natural language processing. Even human intelligence encounters difficulty in mastering idiomatic expressions when learning a foreign language (Miller, 2018). The choice of idioms such as *mala beba* (lit. little baby) and *zlatna koka* (lit. golden goose) for the concept of kindness suggests that the literal meanings of the components were taken into account, with words like ’baby’ and ’golden’ being associated with the concept of being good.

The query was then repeated to ChatGPT, asking it to, in the role of a lexicographer and linguist, find the 10 most relevant idioms in the list of 150 provided Croatian idioms that belong to the semantic fields of madness, conflict, and kindness, i.e. those that are semantically closest to these concepts.

The results matched the manual organization of concepts and idioms in 98% of cases. Three idioms from the concept of kindness were ranked in the top three positions, and nine idioms from the concept of madness were in the top 9 positions. For the concept of conflict, six idioms matched, and ChatGPT did not include the idioms *dolijevati ulje na vatru*, 'further inflame a conflict or disagreement', *prosipati žuč* 'to express bitterness', *ukrstiti koplja* (lit. to cross swords) 'to engage in a conflict' and *stvarati zlu krv* (lit. to create bad blood) while it added the idioms *braniti se rukama i nogama* 'to defend oneself tooth and nail', *digla se kuka i motika* 'to rebel', and *dignuti se na zadnje noge* 'to stand up on hind legs.' In manually assigned concepts, the first one is categorized as avoidance, while the other two are classified as rebellion. The categorization offered by ChatGPT is not necessarily incorrect, as such catego-rization is subject to interpretation, and the usage of idioms greatly depends on context. Conflict typically refers to disagreement, opposition, or tension, while avoidance involves making a deliberate effort to steer clear of conflict or confrontation, which can be associated in some contexts. Also, rebellion implies resistance or opposition against authority or established norms, which can sometimes lead to conflict. In this sense, ChatGPT achieved good results in this experiment.

3.2 Experiment two

In the second experiment, we used ChatGPT. We defined the prompt as follows: based on the list containing 64 idioms and 10 concepts, classify them by meaning into the corresponding fields, essentially matching idioms with con-cepts. ChatGPT categorized the idioms into concepts in the same way as we had previously done manually, resulting in a 100% match. Considering the values provided on both sides and the relatively small number of idioms that needed to be semantically arranged into the proposed concepts, it worked as effectively as human intelligence. In addition, when asked to categorize the 64 idioms based on their meanings and come up with suitable names for each category, it did almost identical work to what we previously did manually. It only separated two idioms from the concept of money into a separate concept of cheapness / low cost: *u bescijenje* 'for a pittance' and *dijeliti šakom i kapom*. However, *dijeliti šakom i kapom* means to give away generously and abundantly, most often money and material things.

3.3 Task one

We conducted previous test experiments to gain insight into the data provided by LLMs. We aimed to identify their strengths and weaknesses. Based on our findings, we have decided to focus our research on the OpenAI's GPT model, as it has shown relatively good results compared to other models. Therefore, the next tasks involve utilizing AI to generate a dataset that lexicographers can use for dictionary creation. As mentioned, there are currently 1,165 idiomatic expressions in the ODCI. Thematic fields were manually identified for 430 entries to establish a dictionary feature that allows users to easily find expressions related to their desired topic or idea through these fields. To ensure accuracy, we wanted to check if the remaining idiomatic expressions can be classified into one of the already manually defined semantic fields.

Prompt used in the experiment:

```
model=”gpt-3.5-turbo”,
messages=[
{”role”: ”system”, ”content”: ”Stavi se u ulogu leksikografa koji stvara novi konceptualno organiziran frazeološki rječnik hrvatskih frazema. Molim te odgovaraj na hrvatskom jeziku.”},
{”role”: ”user”, ”content”: f”Ponuđena je lista s unaprijed definiranim semantičkim poljima.”}
f”Poveži frazem {frazem} s najprikladnijim semantičkim poljem s ponuđenog popisa. Odgovori tako da odabereš samo jedno od ponuđenih semantičkih polja.”
}]
```

To demonstrate the results, we will use the examples of two concepts: communication and knowledge. Using manual classification, we sorted out 19 idioms under the category of communication. In Table 2, we demonstrate how these idioms relate to the results obtained from ChatGPT, which also identified 13 of them as being associated with communication.

Furthermore, under the concept of knowledge, we manually classified the following idioms: *znati što kao vodu piti* ‘to know something like the back of your hand’, *imati u malom prstu što* ‘to have something at your fingertips’ and *isisati iz malog prsta što* ‘to pull something out of thin air, to come up with something effortlessly’. GPT-3.5-turbo classified the idiom *znati što kao* *vodu piti* 'to know something like the back of your hand' under knowledge, while it associated *imati u malom prstu* 'to have something at your fingertips' with the concept of control, and *isisati iz malog prsta što* 'to pull something out of thin air, to come up with something effortlessly' with the concept of easy/difficult. However, GPT-3.5-turbo also classified the idiom *imati do-bar nos* (lit. to have a good nose), which was previously unclassified, under the concept of knowledge, as it means to have the ability or instinct for something (which can include knowledge).

Table 2: Results of Task 1 inquiry using the example of the communication category.

Idioms manually classified into the concept `communication`	ChatGPT-3.5-turbo responses
baciti bubu u uho komu (lit. to plant a bug in someone's ear) 'to make someone suspicious or curious'	`communication`
bacati drvlje i kamenje na koga, što (lit. to throw sticks and stones at someone/something) 'to criticize harshly'	`conflict`
čašica razgovora 'a friendly chat'	`communication`
čupati kliještima iz koga što (lit. to extract something from someone with pliers) 'to forcefully extract information'	`fighting`
pričati Markove konake 'to tell long and boring stories'	`communication`
pričati kao navijen 'to talk incessantly, like a broken record'	`communication`
razgovarati na ravnoj nozi 'to talk on equal terms'	`communication`
reći komu što ga ide 'to tell someone off'	`communication`
reći popu pop, a bobu bob 'to call a spade a spade'	`communication`
reći u lice 'to say to someone's face'	`communication`
šutjeti kao pizda 'to keep silent' (vulgar, lit. to be silent like a cunt)	`communication`
šutjeti kao zaliven 'to be silent as the grave'	`communication`
zatvoriti se u ljušturu 'to withdraw into one's shell'	`unknown`
prosipati pamet 'to dispense wisdom, to pretend to be wise'	`communication`
srati kvake 'to talk nonsense' (vulgar, lit. to shit handles)	`communication`
prenositi se od usta do usta 'to spread by word of mouth'	`communication`
umotati u celofan 'to sugarcoat'	`ingratiation`
obilaziti kao mačak oko vruće kaše 'to beat around the bush'	`avoidance`
lagati u oči komu 'to lie to someone's face'	`fraud`

In addition, GPT-3.5-turbo associated the uncategorized idioms *gurati pod nos komu što* 'to shove something in someone's face (lit. nose), impose something on someone' and the idiom *objaviti na sva zvona* 'to shout it from the rooftops, to announce something to everyone'. Examples of usage for the idiom *to shove something in someone's face* (1 and 2) and for the idiom *to shout it from the rooftops* (3 and 4) found in the ODCI show a context of communication:

1. *If you push your views and principles under his nose on the first date and show him your great intelligence, he will get the impression that you're lecturing him.*
2. *In every argument, he brings up the issues that have been resolved, re-analyzes them, and puts them under the nose.*
3. *After deciding to get engaged, many couples in love don't want to shout it from the rooftops to everyone right away but will keep their sweet secret for some time.*
4. *Don't shout it from the rooftops that you've just received your paycheck, bought new household appliances, or saved a large sum of money, are some of the useful tips that the police have given to citizens.*

Overall, the results offered by GPT-3.5-turbo for Task 1 proved useful for further lexicographical considerations. In other words, they cannot be taken as a finished dataset, but they can assist in providing a comprehensive overview and potential ideas for different categorizations. To improve time efficiency in dictionary creation, a model should have better performance, resulting in fewer mistakes, such as merging *krenuti čijim stopama* 'to follow in someone's foot-steps' with the concept of excitement. This would enable lexicographers to integrate more data with minimal intervention.

3.4 Task two

In the second task, we had the model determine concepts and categorize idioms using the same collection of 1165 idioms. Results showed two already noticed issues regarding the main features of the GPT-3.5-turbo model: the question of well-crafted prompts and the issue of generating always new responses. The first issue suggests that we may have needed to instruct the model to attempt to group a larger number of idioms semantically related to a single concept, rather than constantly offering different concepts.

However, this conflicts with the model's inherent non-deterministic nature, as it always provides a different response to the same prompt. This can be seen in the following example. For a group of idiomatic expressions, the model proposed the following concepts: emotions, emotional reactions, emotional states, and emotional closeness. In Table 3, we present the idioms associated with these concepts. On one hand, the detailed breakdown of the concept of emotions—dividing it into reactions, states, and closeness—can be very useful and aligns with the further subdivision into sub-concepts that we considered in the manual classification. However, if we take into account that a user searches for dictionary entries based on a particular concept, such as happiness, it becomes evident that the field of emotions, even with additional information on reactions, is too abstract and does not fulfill the goal of conceptual organization. The aim is to guide the user, informing them that idioms such as *crven od bijesa* 'red with anger', *kipjeti od bijesa* 'boiling with anger', *ljut kao ris* 'angry as a lynx', *ljut kao vrag* 'angry as the devil', *para ide na uši komu* (lit. steam coming out of someone's ears) 'someone is steam-ing with anger', *pao je mrak na oči komu* (lit. darkness fell over someone's eyes) 'someone saw red', *poludjeti od bijesa* 'go mad with anger', *pozelenjeti od bijesa* 'turn green with anger', and *puknuo je film komu* (lit. someone's film broke) 'someone snapped' are semantically linked to the concept of anger. Similarly, the model assigned the idiom *ne bi ni mrava zgazio* 'wouldn't hurt a fly' the concept of mercy and empathy, and the idiom *duša od čovjeka* (lit. a soul of a man) 'a kind-hearted person' the concept of personality trait. Both are categorized in the manual under the concept of kindness. The con-cepts offered by ChatGPT are not fundamentally wrong in this case, but they do not meet the lexicographer's need to classify all semantically similar id-ioms under the same concept that is general enough to encompass multiple instances, but specific enough to provide users with concrete, usable information. Furthermore, the model assigned some idioms concepts that are semantically linked to the literal meanings of their components. For example, it categorized the idioms *potreban kao kruh* (lit. needed like bread) and *najeo se ludih gljiva tko* (lit. someone ate crazy mushrooms) under the concept of food, although the former means 'urgently needed' and the latter means 'someone went crazy.' For the idiom *mekan kao svila* (lit. soft as silk) 'extremely, very soft' it created the concept of textile properties. Overall, this task did not yield usable data, only a few ideas that can be considered.

Table 3: The concepts proposed by the GPT-3.5-turbo model and associated idioms.

Concept created by the GPT-3.5-turbo	*Associated idiom*
`emotions`	*umrijeti od smijeha* 'die laughing', *tresti se od bijesa* 'shake with anger', *zaljubiti se do ušiju* 'fall head over heels in love', *blagi očaj* 'mild despair', *duša od žene* 'woman with a kind heart', *srce se steže komu* 'someone's heart tightens'
`emotional reaction`	*puknuo je film komu* 'someone snapped, lost it', *dignuti se na stražnje noge* (lit. get up on one's hind legs) 'stand up for oneself', *poludjeti od bijesa* 'to go mad with rage', *rasplakati se kao malo dijete* 'cry like a little child', *plakati kao beba* 'cry like a baby'
`emotional condition`	*nervozan kao pas* 'nervous as a dog', *ljut kao vrag* 'angry as hell', *bijesan kao pas* 'mad as a hornet', *zaljubljen kao tele* 'infatuated, puppy love', *baciti u očaj koga* 'to drive someone to despair'
`emotional closeness`	*zavući se pod kožu komu* 'to get under someone's skin'
`negative emotions`	*proliti žuč* 'to vent one's spleen'

4 CONCLUSION

The purpose of this paper was to examine how large language models respond to inquiries regarding semantic features of multi-word expressions with figurative meanings, specifically idioms. The research was conducted on four large language models: three open-sourced and available for use, Cro-CoV-cseBERT, bcms-bertic, and gpt2-vrabac, and the commercially developed model GPT-3.5-turbo. The results ranged from completely incorrect to very good, with ChatGPT providing the best results. In our concluding discussion, it is important to emphasize that our goal was to obtain usable results in lexicography. However, it should be noted that we queried large language models and a chatbot that uses deep learning to generate human-like responses to natural language queries. It is also important to remember that lexicography is a highly specialized discipline with specific requirements in listing the most typical syntax patterns, selecting collocations and other recurrent phraseological patterns, producing definitions that describe the most important semantic features of a word, and providing examples of usage that reflect the most typical contexts found in the corpus data. These requirements are so stringent that even human intelligence often finds it challenging to meet these expectations and constraints. At present, artificial intelligence has the potential to assist in the creation of dictionaries. To enhance automation for more complex tasks, human intelligence should prioritize the following areas: generating linguistic data for specific languages, particularly those that are small and lack resources, to facilitate the development of robust language models for these languages. Additionally, developing queries that better explain the lexicographic position and needs so that models can produce more effectively applicable results.

5 ACKNOWLEDGMENTS

This work has been fully supported by the University of Rijeka project uniri-mladi-drustv-23-33.

REFERENCES

Babić, K ., Petrović, M., Beliga, S., Martinčić-Ipšić, S., Matešić, M., & Meštrović, A. (2021)Characterisation of COVID-19-Related Tweets in the Croatian Language: Framework Based on the Cro-CoV-cseBERT Model. *Applied Sciences*, *11*(21). https://www.mdpi.com/2076-3417/11/21/10442 doi: 10.3390/app112110442

Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medved, M., Mechura, M., Rychly, P., & Suchomel, V. (2019). Automating dictionary production: a Tagalog-English-Korean dictionary from scratch. , 805-818.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). ELECTRA: Pre-training text encoders as discriminators rather than generators. In *ICLR*. https://openreview.net/pdf?id=r1xMH1BtvB

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

Filipović Petrović, I. (2018). *Kada se sretnu leksikografija i frazeologija: o statusu frazema u rječniku*. Srednja Europa.

Filipović Petrović, I., López Otal, M., & Beliga, S. (2024). Croatian idioms integra- tion: Enhancing the LIdioms multilingual linked idioms dataset. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 4106–4112). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.366

Filipović Petrović, I., & Parizoska, J. (2023). *Frazeološki rječnik hrvatskoga jezika v2*. Hrvatska akademija znanosti i umjetnosti. https://lexonomy.elex.is/#/frazeoloskirjecnikhr

Filipović Petrović, I., & Kocijan, K. (2024). Creating the dataset of Croatian verbal idioms: automatic identification in a corpus and lexicographic implementation. In *Proceedings of the Euralex 2024.* (Accepted for publication)

Filipović Petrović, I., & Parizoska, J. (2019). Konceptualna organizacija frazeoloških rječnika u leksikografiji. *Filologija*, *73*, 27–45.

Fuertes-Olivera, P. A. (2024). Making lexicography sustainable: Using chatgpt and reusing data for lexicographic purposes. *Lexikos*, *34*(1), 123-140. https://lexikos.journals.ac.za/pub/article/view/1883 doi: 10.5788/34-1-1883

Geeraerts, D. (1989). Principles of monolingual lexicography. In F. J. Hausmann (Ed.), *Wörterbücher. ein internationales handbuch zur lexikographie* (Vol. 1, pp. 287–296). Walter de Gruyter.

Hargraves, O. (2018). Information retrieval for lexicographic purposes. In P. A. Fuertes-Olivera (Ed.), *The routledge handbook of lexicography* (pp. 701–714). Routledge.

Jakubíček, M., Kovář, V., & Rychlý, P. (2021). Million-click dictionary: Tools and methods for automatic dictionary drafting and post-editing. In *Book of Abstracts of the 19th EURALEX International Congress* (p. 65-67).

Kocijan, K., Filipović Petrović, I., & Parizoska, J. (2023). Verbal idioms in Croatian: Preparing language data for automatic identification in a corpus. In *International conference Language and language data (CLARC 2023), book of abstracts.* Centar za jezična istraživanja.

Kosem, I., Gantar, P., Logar, N., & Krek, S. (2014). Automation of lexicographic work using general and specialized corpora: Two case studies. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the 16th EURALEX International Congress* (p. 355-364). EURAC research.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences*

Communications, *10*(704). doi: 10.1057/s41599-023-02119-6

Ljubešić, N., & Klubička, F. (2016). *Croatian web corpus hrWaC 2.1.* Slovenian language resource repository CLARIN.SIhttp://hdl.handle.net/11356/1064

Ljubešić, N., & Kuzman, T. (2024). CLASSLA-web: Comparable web corpora of South Slavic languages enriched with linguistic and genre annotation. In N. Cal- zolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3271–3282). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.291

Ljubešić, N., & Lauc, D. (2021). BERTić - the transformer language model for Bosnian, Croatian, Montenegrin and Serbian. In B. Babych et al. (Eds.), *Pro- ceedings of the 8th Workshop on Balto-Slavic Natural Language Processing* (pp. 37–42). Association for Computational Linguistics. https://aclanthology.org/2021.bsnlp-1.5

Ljubešić, N., Rupnik, P., & Kuzman, T. (2024). *Croatian web corpus CLASSLA-web.hr 1.0.* Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1929

McArthur, T. (1986). *Worlds of reference: Lexicography, learning, and language from the clay tablet to the computer*. Cambridge University Press.

Miller, J. (2018). Research in the pipeline: where lexicography and phraseology meet. *Lexicography ASIALEX*, *5*(1), 23–33. doi: 10.1007/s40607-018-0044-z

Moussallem, D., Sherif, M. A., Esteves, D., Zampieri, M., & Ngonga Ngomo, A.-C. (2018, May). LIdioms: A Multilingual Linked Idioms Data Set. In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018).* European Language Resources Association (ELRA). https://aclanthology.org/L18-1392

Perak, B., Beliga, S., & Meštrović, A. (2024). Incorporating dialect understand- ing into LLM using RAG and prompt engineering techniques for causal commonsense reasoning. In Y. Scherrer, T. Jauhiainen, N. Ljubešić, M. Zampieri, P. Nakov, & J. Tiedemann (Eds.), *Proceedings of the 11th Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)* (pp. 220–229). ACL. https://aclanthology.org/2024.vardial-1.19 doi: 10.18653/v1/2024.vardial-1.19

Rehm, G., & Way, A. (2023). *European language equality: A strategic agenda for digital language equality*. Springer Nature. https://doi.org/10.1007/978-3-031-28819-7 doi: 10.1007/978-3-031-28819-7

Rundell, M. (2023). Automating the creation of dictionaries: Are we nearly there? In *Proceedings of the 16th international conference of the asian association for* *lexicography: Lexicography (asialex 2023 proceedings)* (pp. 1–9). Yonsei University. (22–24 June 2023)

Tran, H. T. H., Podpečan, V., Jemec Tomazin, M., & Pollak, S. (2023). Definition Extraction for Slovene: Patterns, Transformer Classifiers and C hatGPT. In M. Medveď, M. Měchuar, C. Tiberius, I. Kosem, J. Kallas, & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century* (pp. 19–38). Lexical Computing.

Ulčar, M., & Robnik-Šikonja, M. (2020). FinEst BERT and CroSloEngual BERT: Less Is More in Multilingual Models. In *Text, Speech, and Dialogue: 23rd International Conference, TSD 2020, Brno, Czech Republic, September 8–11, 2020, Proceedings* (p. 104–111). Springer-Verlag. https://doi.org/10.1007/978-3-030-58323-1_11 doi: 10.1007/978-3-030-58323-1_11

Vossen, P. (2022). Chatgpt is a waste of time. *VU Magazine*. Retrieved from https://vumagazine.nl/professor-piek-vossen-chatgpt-is-a-waste-of-time?lang=en

Škorić, M. (2024). Novi jezički modeli za srpski jezik. *Infoteka*, *24*. https://arxiv.org/abs/2402.14379

VELIKI JEZIKOVNI MODELI PODPIRAJO LEKSIKOGRAFIJO: KONCEPTUALNA ORGANIZACIJA HRVAŠKIH IDIOMOV

V tem članku opisujemo kako veliki jezikovni modeli odgovarjajo na poizvedbe o semantičnih značilnostih idiomatskih izrazov v hrvaščini. Natančneje, ustvarili smo poizvedbe za štiri velike jezikovne modele z uporabo vzorca 430 idiomov iz prosto dostopnega *Spletnega slovarja hrvaških idiomov*. Ti idiomi so bili prej kategorizirani v 65 konceptov ali semantičnih kategorij. Ker so to delo ročno opravili jezikoslovci in leksikografi, smo želeli raziskati kakovost in natančnost rezultatov, pridobljenih s sistemi, ki temeljijo na umetni inteligenci, in jih primerjati z rezultati, pridobljenimi s človeško inteligenco. Cilj je bil oceniti, ali so odgovori primerni za leksikografsko obdelavo in v kolikšni meri jih lahko leksikografi uporabljajo, morda kot zanesljivo orodje za samodejno ustvarjanje konceptualne organizacije idiomov.

Keywords: veliki jezikovni modeli, idiomi, semantična podobnost, konceptualna organizacija v leksikografiji