AI- and Corpus-Based Strategies for Identifying Phraseme Constructions: A Pilot Study on Croatian Repetitive Constructions

Slobodan Beliga[1,2], Ivana Filipović Petrović[3]

[1]Faculty of Informatics and Digital Technologies, University of Rijeka, Radmile Matejčić 2, 51000 Rijeka, Croatia

[2]Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Trg braće Mažuranića 10, 51000 Rijeka, Croatia

[3]Croatian Academy of Sciences and Arts, Trg Nikole Šubića Zrinskog 11, 10000 Zagreb, Croatia

E-mail: sbeliga@inf.uniri.hr, ifilipovic@hazu.hr

Abstract

The paper introduces a hybrid methodology for cross-linguistic identification of phraseme constructions, developed within the scope of a pilot study on Croatian repetitive constructions. The study explores how artificial intelligence and corpus technologies can be systematically combined to uncover functionally equivalent patterns across languages. The proposed strategy rests on three interdependent layers: (1) the AI layer, which harnesses large language models to generate candidate constructions, paraphrases, and corpus query formulations; (2) the corpus layer, which provides empirical validation through frequency data, authentic usage, and syntactic patterns; (3) and the human expert layer, which supervises prompt engineering, interprets outputs, and ensures linguistic adequacy. These layers operate in an iterative workflow, enabling dynamic interaction between computational and expert insights. The methodology is exemplified through the analysis of the German construction `X über X 'X after X'`, for which the Croatian equivalent `X za X-om` (e.g., *dan za danom* 'day after day') is identified as structurally and semantically appropriate. The study compares outputs of two LLMs (GPT-4o and o3), revealing performance differences in idiomatic sensitivity. It also demonstrates how LLMs can assist in filtering corpus concordances to identify phraseologically valid examples. The study highlights both the strengths (e.g., scalability, reduced expert workload) and limitations (e.g., LLMs' sensitivity to prompt design and formal syntax) of the approach. It concludes that this layered strategy offers a viable path toward the semi-automatic processing of additional constructions and the development of multilingual phraseological resources.

Keywords: Repetitive phraseme constructions; Corpus Query Language; Large Language Model; AI-assisted lexicography; Human-in-the-loop NLP

1. Introduction

Phraseme constructions have become an increasingly prominent topic in recent phraseological research, particularly within the framework of construction grammar (Goldberg, 1995, 2019). Characterized by a combination of fixed elements and open slots that can be filled in context, they require access to authentic language data for proper analysis. Large computational corpora have therefore emerged as essential tools in investigating these constructions across real-world usage. Recent studies further suggest that artificial

intelligence (AI) tools can complement corpus-based approaches by supporting specific linguistic tasks (Davies, 2025).

These perspectives gain particular relevance in the context of the COST Action PhraConRep: *A Multilingual Repository of Phraseme Constructions in Central and Eastern European Languages* [1] (CA22115), which focuses on building a multilingual repository of phraseme constructions using German as the source language and 14 target languages, including Croatian. The starting point for this effort is a curated list of 339 constructions available in the PhraConRep[2] repository. Project members, philologists with expertise in both the target language and German, are working to identify equivalents in published literary translations, which serve as initial entry points for deeper investigation of how such constructions are realized in the target languages.

Phraseology is known to operate on both a global and local level: while many phraseological patterns are cross-linguistically widespread, each language also reflects its own specific linguo-cultural realizations. The same applies to phraseme constructions, which, although structurally comparable across languages, often display language-specific variation in their open slots. These differences may reflect cultural references, regional concepts, or grammatical and syntactic features unique to each language.

In the case of Croatian, the repository currently contains equivalents of the German construction following the repetitive pattern `X über X` (e.g., *Fehler über Fehler*, 'mistakes after mistakes', *Tage über Tage* 'day after day'), that were taken from existing literary translations. These individual translation solutions include, for instance, *Briefe über Briefe* ('a flood of letters'), which was translated as *sila pisama*, a semantically adequate but structurally different equivalent that does not preserve the repetitive pattern. Another example, *Gründe über Gründe*, was translated into Croatian as *razlozi nad razlozima*, which, despite using a similar preposition (*über/nad* 'above'), alters the meaning: instead of expressing iteration ('many reasons'), it conveys a hierarchical relation ('the most important reasons'). A closer match would be *razlozi za razlozima* ('reason after reason'), which maintains the iterative semantics. However, corpus data show that this construction is extremely rare in Croatian and typically appears in the singular form.

These examples illustrate two key challenges in cross-linguistic research on phraseme constructions. First, semantic equivalence does not always align with structural correspondence, which is an observation that is both well established and widely expected within the field of translation. Second, a translated example proposed by a particular translator should be treated merely as a preliminary hint, not as a definitive solution. Even when a structurally similar equivalent exists, it may be extremely infrequent in actual usage, as corpus data show.

To address these issues, this study proposes a methodology that combines AI tools and corpus-based methods for identifying and verifying phraseme constructions across languages. As a pilot case, we focus on the German construction `X über X` and examine its functional and structural equivalents in Croatian, aiming to test a scalable strategy for broader application to other constructions from the PhraConRep database. Specifically, the following research objectives guide the pilot study:

[1] COST Action CA22115 – PhraConRep: https://www.phraconrep.com/

[2] PhraConRep (2024). Phrasemkonstruktionen Wörterbuch (PhKW). https://github.com/PhKW/PhKWB/tree/main/Artikel

62. To evaluate whether Large Language Models (LLMs) can autonomously propose plausible repetitive phraseme constructions in Croatian based on analogy with the German `X über X`.
63. To assess the ability of LLMs to generate Corpus Query Language (CQL) expressions for identifying such constructions in Croatian corpora.
64. To examine a human-in-the-loop strategy that combines LLM-generated suggestions with corpus data, aiming to validate or refine candidate constructions based on frequency and context.

This paper is organized as follows: it begins with the theoretical grounding (Section 2), followed by an exploration of corpora and AI tools in phraseological research (Section 3). Section 4 outlines the data and methods, Section 5 details the methodological steps, Section 6 presents the results, and Sections 7 and 8 provide discussion and final reflections.

2. Theoretical Background on Phraseme Constructions

Phraseme constructions are idiomatic multi-word units that lie at the intersection of the lexicon and grammar. They typically consist of one or more synsemantic anchor elements – lexically fixed components – and one or more open slots that are contextually filled with lexical items (Pavlova et al., 2022: 182). These open positions allow for variation, while the anchor elements remain stable, thus combining formal rigidity with contextual flexibility.

These constructions represent cases in which one part (the anchors) is obligatory and fixed, while the remaining parts (the slots) are variable. The range of possible fillers for these open positions can be wide, but the lexical items that appear in these slots tend to form a frequency-based continuum. This continuum can be semantically categorized, and from this categorization, it is possible to derive pragmatic functions. While constructions of this type exhibit idiomatic meaning as wholes, their internal structure reflects a combination of fixed lexical elements and semantically constrained variation (Dobrovol'skij, 2011, 2022).

The development of construction grammar has provided a suitable theoretical framework for analyzing these patterns, especially through corpus-linguistic methods. Construction grammar views language as a network of form–meaning pairings shaped by usage, offering tools to describe both structural regularity and lexical variation. Key overviews of this intersection include Ziem (2018), with recent elaborations by (Dobrovol'skij, 2018), (Mellado Blanco, 2022, 2023) and (Piunno, 2022), who refer to such patterns as Phrasem-Konstruktionen, semi-schematic constructions or partially lexically filled units. While terminology may vary, these approaches converge on the idea of combining structural regularity with lexical variability.

To illustrate, consider the following Croatian phraseme constructions:

- `ako X, onda Y`
  Ako je to umjetnost, onda sam ja Mojsije. 'If that is art, then I am Moses.'
- `jednom X, uvijek X`
  Jednom učitelj, uvijek učitelj. 'Once a teacher, always a teacher.'
- `X koji/koja/to nije`
  Komedija koja to nije, mir koji to nije. 'A comedy that is not really a comedy, peace that is not really peace.'

The open slots in these constructions may be filled with a wide array of lexical items, often producing humorous, ironic, or hyperbolic effects. However, the choice of fillers is not entirely arbitrary. Rather, it is governed by morphosyntactic and semantic constraints, and the patterns of variation can be classified, analyzed, and quantified through corpus-based research (Steyer, 2013). Large corpora thus make it possible to investigate the full range of formal and semantic variants of these constructions across different languages.

Phraseme constructions frequently perform an evaluative pragmatic function. They tend to be highly expressive, making them especially suitable for use in spoken discourse (Pavlova, 2024; Mellado Blanco, 2022). This evaluative potential can be seen as a key factor in the phraseologization process, positioning such constructions within the category of phrasemes with a distinct communicative-pragmatic role.

3. Leveraging Corpora and LLMs in Phraseological and Lexicographic Research

Corpora have long been central to linguistic and lexicographic research, offering reliable, authentic insights into language use. They provide data on frequency, collocations, variation across genres and time, and, crucially, on how multiword expressions such as phraseme constructions and idioms function in real contexts. Tools like Sketch Engine and similar platforms have made it possible to extract and visualize such information efficiently, empowering researchers to base their analyses on observable patterns in usage rather than introspective judgments.

The traditional strengths of corpora – empirical grounding and transparency – are now being complemented by LLMs, which bring new capacities to interpret, group, and explain linguistic data. As Mark Davies (2025) notes, the integration of LLMs into corpus interfaces (as in English-Corpora.org) enables users to send collocates, concordances, and frequency data directly to models like GPT[3] or Gemini[4], which then generate semantic groupings, paraphrases, or functional descriptions.

Recent studies confirm both the promise and current limitations of LLMs in phraseological and lexicographic research across diverse languages and tasks. AbuMandour (2024) propose the AI-Model Triangulation Approach (AMTA), combining multiple AI tools (e.g., ChatGPT, Gemini) with prompt engineering and corpus data to generate bilingual Arabic–English dictionary entries in technical domains. Their results show strong potential, particularly in semantic precision and contextual relevance, while minimizing expert intervention. Similarly, Li et al. (2024) introduce IDIOMKB, a multilingual knowledge base that, through a retrieval-augmented strategy, significantly improves idiom translation performance in smaller models (e.g., BLOOMZ, Alpaca), especially in recovering figurative meaning. In contrast, Chen et al. (2024) demonstrate that while ChatGPT-4(o) can generate plausible French–Chinese phraseme equivalents, performance is highly sensitive to prompt formulation and contextual framing. Shalevska et al. (2025) report that ChatGPT, Gemini, and DeepSeek struggle to preserve stylistic features such as rhyme and metaphorical nuance when translating Macedonian idioms, often producing overly literal renderings. Similarly, Sørensen & Nimb (2025) show that LLMs frequently fail to detect culturally embedded figurative meanings in Danish idioms and are prone to accepting

[3] Generative Pre-trained Transformer - developed by OpenAI.
[4] Sparse mixture-of-experts (MoE) transformers - developed by Google.

misleading definitions when convincingly phrased. For Slovenian, Gantar (2024) observe that ChatGPT provided adequate or superior dictionary definitions for over half of tested idioms, but struggled with example generation and polysemy. In Croatian studies, Beliga & Filipović Petrović (2024) report 68% accuracy for GPT-3.5 in classifying idioms into predefined semantic domains, while follow-up work by Filipović-Petrović & Beliga (2024); Beliga & Filipović Petrović (2025) shows that GPT-4o outperformed its predecessor in unsupervised idiom clustering, producing coherent categories such as love and affection or anger and frustration, contingent on carefully crafted prompts.

Taken together, the reviewed studies underscore both the expanding role and the persistent limitations of LLMs in lexicographic and phraseological research. While LLMs demonstrate considerable potential in recognizing figurative meaning and generating idiomatic equivalents, they continue to face challenges in producing functionally and culturally appropriate interpretations across languages. Their performance remains highly sensitive to prompt formulation, and their context modeling in non-literal discourse is often inconsistent. These observations reaffirm the complementary relationship between corpora and LLMs. Corpora offer empirical grounding, frequency data, and access to authentic usage patterns, which are essential for linguistic transparency and verification. In contrast, LLMs contribute through semantic abstraction, constructional generalization, and interpretive flexibility. A methodological synthesis of both resources—augmented by expert input and iterative evaluation—thus enables a more robust and scalable approach to the analysis of multiword expressions, as advanced in this study.

4. Data and Methods

This section outlines the linguistic and technical resources that underpin the proposed strategy, including the construction type examined, the dataset used, and the corpus and AI tools employed.

4.1 Data Sources

Corpus: CLASSLA-web.hr. This study uses CLASSLA-web.hr 1.0, the largest publicly available linguistically annotated Croatian web corpus, developed by Ljubešić et al. (2024) within the South Slavic CLASSLA-web initiative. It comprises ~2.58 billion tokens and 5.4 million texts, derived from the MaCoCu-hr 2.0 crawl (2021–2022), which targeted the .hr domain and additional Croatian-language sources (Ljubešić & Kuzman, 2024). The corpus was filtered to exclude non-Croatian and very short texts, then annotated via the CLASSLA-Stanza pipeline (tokenization, lemmatization, morphosyntactic tagging). Genre labels were automatically assigned using the X-GENRE classifier (XLM-RoBERTa), yielding ten categories (e.g., *News*, *Opinion*, *Legal*). Provided in vertical format, the corpus is CWB/Sketch Engine–compatible and enriched with multi-level metadata (e.g., text ID, URL, genre, paragraph quality via jusText).

Source Dataset: PhraConRep. The PhraConRep repository[5] contains 339 phraseme constructions with German as the source language (PhKW, 2024). Each entry follows a uniform lexicographic structure, covering: *lemma, variants, meaning, examples, morphology, syntax, usage, and function*, which ensures comparability and facilitates integration into

[5] https://github.com/PhKW/PhKWB/tree/main/Artikel

fourteen Central and Eastern European target languages. All entries are exemplified by authentic German usage excerpts drawn from literary texts, journalistic discourse, websites, and representative corpora, thereby grounding the resource in real-world language data. Although multilingual expansion is still in progress, the repository's consistent schema supports contrastive analysis and tracking of semantic, syntactic, and pragmatic correspondences across languages.

Repetitive Constructions. This study isolates repetitive phraseme constructions as a distinct formal-functional category. These constructions exhibit fixed structural patterns (e.g. `once X, always X; X as X; X over X; from X to X; X stays X`) and serve pragmatic functions such as intensification, contrast, iteration, evaluation, or irony. Repetition typically involves identical or near-identical elements and adds rhetorical emphasis, expressivity, or humorous effect in discourse (cf. Hohenhaus, 2004; Horn, 2018). Repetition is approached here both as a linguistic feature and a methodological strategy. It enables the construction of idiomatic meaning and offers a heuristic for CQL-based corpus queries.

No.	Construction Pattern	German Examples with English translations
1	`X ist / bleibt X` `'X is / remains X'`	*Krieg ist Krieg* 'war is war' *Kind bleibt Kind* 'a child remains a child'
2	`XX / X-X`	*Kaffee-Kaffee* (lit. coffee-coffee) 'real coffee' *kochen-kochen* (lit. cooking-cooking) 'proper cooking'
3	`X über X` `'X upon X / X over X'`	*Fehler über Fehler* 'mistakes upon mistakes' *Tage über Tage* 'days upon days'
4	`X der X / X aller X` `'X of all Xs / the X of Xs'`	*das Meer der Meere* 'the sea of all seas' *die Besten der Besten* 'the best of the best'
5	`einmal X, immer X` `'once a(n) X, always a(n) X'`	*einmal Maler, immer Maler* 'once a painter, always a painter'
6	`es gibt X und X` `'there are Xs and Xs'`	*es gibt Kinder und Kinder* 'there are children and there are children'
7	`von X zu X` `'from X to X'`	*von Tag zu Tag* 'from day to day'

Table 1: Data from the PhKW repository

From a list of 339 phraseme constructions, seven repetitive patterns were manually selected as representative of this constructional type (see Table 1). Among them, the construction `X über X 'X upon X / X over X'` was chosen for in-depth analysis. Treated as a paradigmatic instance, it enables the clear and systematic exposition of the proposed methodology at the interface of pattern recognition and corpus-informed interpretation.

Selected for its structural regularity, semantic versatility, and cross-linguistic relevance, this construction serves as an analytic prototype for integrating symbolic querying (via CQL) with LLM-assisted semantic modelling. The approach prioritizes methodological transparency over exhaustive coverage, providing a foundation for scalable application to additional constructions in future research.

4.2 Tools and LLMs

noSketch Engine and CQL. Corpus queries were conducted using the noSketch Engine, an open-source interface derived from the commercial Sketch Engine and accessed via CLARIN.SI, to search the CLASSLA-web.hr corpus using CQL. This setup enabled the extraction of complex syntactic patterns, such as phraseme constructions, by combining fixed lexical anchors with variable slots in structured queries. The engine supports KWIC browsing, collocation analysis, frequency profiling, and genre filtering, making it suitable for advanced linguistic research. It allows for precise retrieval of both fixed and variable constructions, offering contextualized concordances, frequency data, and document-level metadata. The interface also facilitated the execution and evaluation of LLM-generated CQL queries, enabling syntactic validation and semantic plausibility checks during candidate extraction. CQL is a formal query language specifically designed for structured corpus searches, enabling users to specify linguistic constraints at the level of tokens, parts of speech, lemmas, and syntactic patterns. It allows researchers to search for complex grammatical or lexical patterns and to define query parameters that are not supported by predefined filtering options in conventional corpus interfaces.

LLMs. This study employs two advanced LLMs developed by OpenAI: GPT-4o and the o3 model. **GPT-4o** is a multilingual and multimodal generative pre-trained transformer (GPT) capable of real-time processing and generation of text, audio, and visual inputs (OpenAI, 2024). It has demonstrated state-of-the-art performance across a range of benchmarks, including automatic speech recognition, machine translation, and multimodal reasoning, and supports over 50 languages, including Croatian. In this study, GPT-4o was primarily employed for identifying structural patterns in German phraseme constructions and generating semantically and syntactically plausible Croatian equivalents, including paraphrastic variants and contextual interpretations guided by prompt engineering. In contrast, the **o3** model is a reflective transformer optimized for step-by-step reasoning through reinforcement learning (OpenAI, 2025). Specifically, the model incorporates a chain-of-thought mechanism developed via reinforcement learning strategies that reward internal deliberation prior to response generation. This allows o3 to simulate intermediate reasoning steps, thereby improving consistency and precision in tasks with formal syntactic constraints, such as the formulation of CQL expressions. All LLM-based experiments were conducted between April 1 and April 10, 2025, using default model settings, including the standard temperature parameter. The selection of GPT-4o and o3 was informed by their respective strengths: GPT-4o's ability to generalize across languages and interpret phraseological semantics, and o3's enhanced performance in rule-based syntax generation and formal reasoning.

5. Methodology

This study proposes a hybrid, AI-assisted corpus-based methodology for identifying functionally comparable phraseme constructions across languages. The core question

addressed is: how can a linguist identify a Croatian construction that corresponds in meaning, usage, or structure to a given German model (`X über X`)? The objective is to find a functionally comparable construction that demonstrates similar phraseological behavior within the target language. The step-by-step strategy used in this study is outlined in Figure 1.

The methodological approach to phraseme construction strategies proposed in this study rests on three interdependent pillars: (1) the AI layer, (2) the corpus layer, and (3) the human expert layer. The **AI layer** encompasses the use of LLMs to support semantic abstraction, paraphrase generation, generation of constructional patterns, formulation of CQL queries, and validation (filtering) of constructions that are credible equivalents in another language. The **corpus layer** provides screening of candidates in a real corpus, empirical grounding through frequency data, authentic usage patterns, and genre-specific variation, enabling data-driven identification and contextual evaluation of constructions. The **human expert layer** ensures interpretive oversight, guiding prompt engineering, query refinement, supervising the interpretation of model outputs, and the validation of model outputs through linguistic and phraseological expertise. The process of constructing phraseme candidates circulates iteratively across all three layers to ensure the linguistic validity and contextual adequacy of the resulting constructions. As visualized in the diagram (see Figure 1), transitions between layers are represented by dashed arrows, indicating the transfer of linguistic data or metadata across components of the workflow.

Viewed through the lens of linguistic engineering, the proposed methodology unfolds as a multi-step, human-AI collaborative process. It begins with a lexicographically informed selection stage, wherein a human expert manually identifies instances of the target repetitive construction within the PhraConRep repository (see Figure 1, step ①). These examples, which are drawn from authentic sources and not accompanied by metalinguistic annotations, are then submitted to a LLM via prompt engineering (step ②), using *Prompt A* (see Appendix). The LLM is tasked with interpreting the input both semantically and syntactically, generating Croatian candidate equivalents of the original German phraseological constructions. Simultaneously, it provides contextualized semantic interpretations and proposes a formalized structural pattern that generalizes over the presented constructions (step ③). Subsequently, the LLM generates a corresponding CQL query designed to retrieve additional instances of the proposed constructional pattern from an external corpus (step ④). This CQL query is then reviewed, and if necessary, revised by the human expert (⑤), before being executed within the corpus analysis system (i.e., noSketch Engine) on the linguistically annotated Croatian corpus (step ⑥). Based on the resulting raw concordance list, a second round of prompt engineering (step ⑦) is initiated using *Prompt B* (see Appendix). Here, the LLM is instructed to filter the concordance data by retaining only those that qualify as genuine phraseological constructions based on contextual coherence and idiomaticity (step ⑧). The filtered output is then further reviewed and refined by the human expert, who conducts a final evaluative step (step ⑨), yielding a validated set of Croatian constructions (step ⑩) that constitute reliable equivalents of the original German phraseological forms.

This elaborated AI-assisted lexicographic hybrid approach ensures that each candidate construction is triangulated across layers, enhancing reliability and enabling a high degree of precision in cross-linguistic phraseme identification. It also illustrates the indispensable

role of expert oversight and the dynamic interplay between computational and empirical linguistic methods within a contemporary paradigm of phraseological research.

The following sections present a simplified three-phase representation of the implementation of the methodology described above. These phases correspond respectively to Figure 1: Phase 1 covers steps ①–③, Phase 2 corresponds to steps ④–⑥, and Phase 3 covers steps ⑦–⑩.

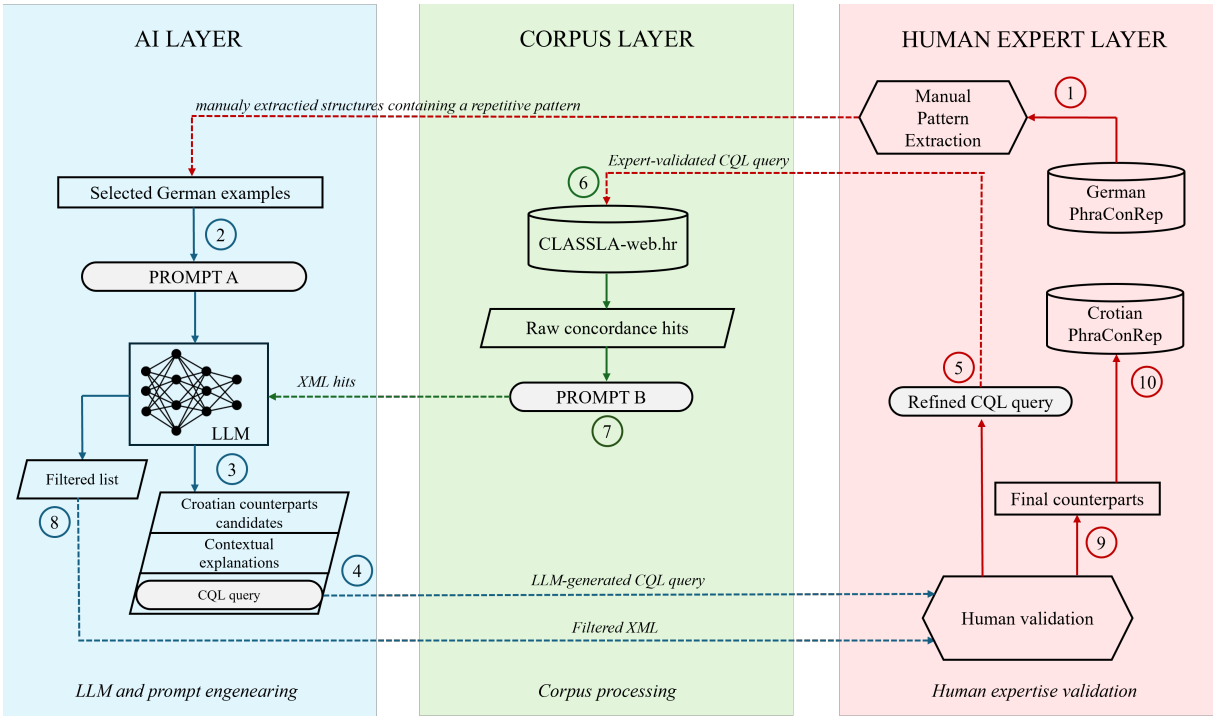

Figure 1: Proposed strategy, i.e. experimental workflow

5.1 Phase 1: AI-Assisted Interpretation and Translation of German Examples

This phase begins with the manual selection of examples of the German repetitive construction X über X from the PhraConRep repository (step ①). These examples are submitted to an LLM using Prompt A (step ②), without any accompanying lexicographic metadata. This decision rests on the assumption that authentic contextual information will generate the most reliable output, whereas metadata such as definitions or usage notes reflect the lexicographer's metalinguistic interpretation rather than actual language use. The LLM is first prompted to interpret the German sentences both semantically and syntactically. By relying solely on contextualized usage, the model is encouraged to infer both meaning and constructional patterns in a way that mirrors human language processing. A follow-up prompt then instructs it to propose comparable constructions in Croatian. This approach relies on in-context learning, combining implicit few-shot prompting with prompt chaining to elicit semantic generalization and cross-linguistic mapping (step ③). Thereafter, the LLM ultimately generates a final CQL query (step ④), based on the entire preceding context regarding the construction type from both German and the proposed Croatian constructions. The output from this phase (CQL query) serves as the input for corpus-based analysis.

5.2 Phase 2: Formulating CQL Queries and Performing Corpus Searches

Given that phraseme constructions are highly context-sensitive and often involve semantically constrained variability, corpus-based analysis is essential for testing and validating proposed equivalents. The input for data retrieval of these proposed constructional patterns within the corpus layer is enabled by an LLM-generated CQL query which serves as a formalized transfer of the distilled knowledge and insights derived from the AI layer. Therefore, the formulation of this CQL query, based on the AI layer's insights, aims to strike a balance between specificity and coverage: overly broad queries produce irrelevant data, while overly narrow ones may miss productive variants. The query is reviewed and, if necessary, refined by the expert linguist (step ⑤), then executed within the noSketch Engine interface using the CLASSLA corpus of Croatian (step ⑥). This corpus-based step enables the empirical verification of proposed constructions.

5.3 Phase 3: AI-Assisted Interpretation of Retrieved Concordances

In the final phase, the raw concordance lines retrieved from the corpus are submitted to the LLM again using Prompt B (step ⑦). The LLM is tasked with identifying which of the concordances truly qualify as phraseme constructions, based on contextual coherence, idiomaticity, and communicative function (step ⑧). The filtered output is then subject to expert validation (step ⑨), resulting in a finalized and linguistically verified list of Croatian equivalents (step ⑩). This completes the iterative cycle across AI, corpus, and expert layers, culminating in a robust set of phraseme construction candidates aligned with the original German model.

6. Results and Evaluation

This section presents the results of the pilot analysis, focusing on the German phraseme construction `X über X 'X after X / X upon X'` (e.g., *Fehler über Fehler* 'mistakes after mistakes', *Tage über Tage* 'days upon days') and its potential Croatian equivalents.

6.1 Data retrieved from LLMs

In the first prompt (*A*), the GPT-4o model was initially asked to interpret several authentic German sentences that exemplify a specific repeated phraseological pattern. Prompt A was formulated as follows: *Could you explain what these German sentences mean? List of selected sentences: [... 32 original sentences from the PhraConRep repository ...]. Based on the entire set of sentences, identify and describe the repeated phraseological construction that they instantiate*. In response, GPT-4o correctly identified the pattern `X über X 'X upon X / X over X'` as a repetitive structure that expresses abundance, accumulation, or intensification through lexical repetition, having generalized this construction from the entire set of 32 sentences rather than from individual sentences. The semantic analysis emphasized its rhetorical and evaluative function, highlighting large quantities or overwhelming presence of the repeated entity. Furthermore, expert linguistic assessment confirmed that the model provided accurate and contextually appropriate explanations for all 32 sentences, with no errors observed in its semantic analyses. In addition to the structural and semantic analysis, the model also provided brief contextual interpretations of the examples, comments on the evaluative function of the constructions, and occasionally remarks on register. Overall,

this first step demonstrates the model's strong ability to abstract a shared phraseological construction and to provide accurate semantic analyses, offering a solid basis for subsequent interpretation and the search for Croatian equivalents.

In the follow-up prompt, the model was asked to provide Croatian equivalents for the target construction. GPT-4o proposed two structural candidates:

- `X preko X 'X over X'` (e.g., *čuda preko čuda* 'miracles over miracles')
- `X na X 'X on X'` (e.g., *uvrede na uvrede* 'insults piled on insults')

While `X na X 'X on X'` may occasionally be acceptable in certain combinations, such as *uvrede na uvrede* ('insults piled on insults'), it is not the dominant or most idiomatic form in Croatian. The other suggestion, `X preko X 'X over X'`, was evaluated as implausible in the intended meaning and is not a typical or natural expression in native usage.

In contrast, the preposition *za* 'for' (as in *pitanje za pitanjem* 'question after question' [lit. question for question], *račun za računom* 'bill after bill' [lit. bill for bill], *uvreda za uvredom* 'insult after insult' [lit. insult for insult]) is the most pragmatically and syntactically appropriate, and is, in fact, the productive form used in Croatian to express sequential or accumulative reduplication. While the model did include some correct examples in its answers, such as *pitanje za pitanjem* 'question after question' and *račun za računom* 'bill after bill', it did not prioritize `X za X-om (lit. X for X) 'X after X'` as the main pattern, instead presenting it alongside less idiomatic or less frequent alternatives. This may reflect a source-language bias, with the model overextending the German structure into Croatian without sufficient regard for target-language constraints. Such a bias is particularly likely when data for the target language is less thoroughly represented in the model's training set.

To better understand the nature and extent of these shortcomings, the model's output was subjected to a targeted expert evaluation. The evaluation of GPT-4o's output was conducted as a qualitative expert analysis due to the limited sample size. A phraseology specialist assessed the model's output along three dimensions: (1) structural validity of the proposed construction pattern, (2) semantic adequacy of its interpreted function (e.g., expression of quantity, accumulation, intensification, and evaluative meaning), and (3) contextual appropriateness of the accompanying explanations, including the provision of idiomatic examples reflecting authentic usage. Each category was rated on a three-point scale (1 – *inaccurate*, 2 – *partially accurate*, 3 – *fully accurate*). GPT-4o's suggestions (`X preko X` and `X na X`) were rated as partially accurate (2): although structurally possible, they were neither dominant nor frequent in Croatian usage, and several examples provided (uvrede na uvrede, čuda preko čuda) were judged atypical in authentic discourse. In contrast, the o3 model was rated as fully accurate (3) across all categories, successfully identifying the dominant pattern `X za X-om` and providing idiomatic examples together with contextually appropriate explanations. This systematic assessment makes it possible to compare the two models more directly and to understand where GPT-4o's limitations occur and how the o3 model improves on them.

After receiving suboptimal output from GPT-4o for the German `X über X` construction, the experiment was repeated with o3 model, which resulted in a substantial improvement in both linguistic accuracy and appropriateness. Unlike GPT-4o, which had over-relied

on literal mappings from German, the o3 model correctly identified the idiomatic and grammatically accurate Croatian pattern: `X za X-om (lit. 'X for X')`, with the first noun in the accusative singular and the second in the instrumental singular. The model's output suggested that in this construction, the preposition *za* (lit. for) loses its typical directional or purposive meaning and instead signals temporal or cumulative succession, effectively functioning as 'after' in English.

In addition, the o3 model offered a comparative overview of related structures, helping to clarify nuances in Croatian usage. It noted a variant with *na* (lit. on) (e.g., *dug na dug* 'debt upon debt', *račun na račun* 'bill upon bill'), in which both nouns remain in the same case and the preposition *na* contributes a sense of accumulation or piling. According to the o3 model, although less idiomatic, this variant can occur in specific lexical pairings.

Finally, the model mentioned occasional use of the preposition *preko* ('over, across') in constructions such as *pitanja preko pitanja* 'questions over questions' or *problemi preko problema* 'problems over problems'. According to the model, these forms mirror the German *über* more literally, but they are considerably rarer in Croatian and generally fall outside the conventional idiomatic repertoire.

Overall, the output from o3 was more accurate in capturing idiomatic usage and more linguistically informative, offering relevant structural distinctions and highlighting context-dependent variation. This supports a broader methodological point: AI tools should be tested across different model versions (and, where relevant, across different model families) as linguistic performance may vary significantly between them.

LLM	**Pattern in Croatian**	**Croatian Examples**
GPT-4o	`X na X` `(lit. X on X) 'X upon X'`	*uvrede na uvrede 'insults piled on insults* *dugovi na dugove 'debt upon debts'*
	`X preko X` `(lit. X over X)`	*čuda preko čuda 'miracles over miracles'* *razloga preko razloga 'reasons over reasons'*
o3	`X za X` `(lit. X for X) 'X after X'`	*pitanje za pitanjem 'question after question'* *greška za greškom 'mistake after mistake'* *nesporazum za nesporazumom* *'misunderstanding after misunderstanding'*

Table 2: LLM-generated Croatian equivalents

In sum, the LLM generated a linguistically insightful set of Croatian candidate expressions (Table 2). However, the quality and reliability of the data varied depending on factors such as cross-linguistic interference and the specific model version used. These findings support the conclusion that human input and critical evaluation remain essential when leveraging LLM-generated data in linguistic research.

Finally, the LLM generated a CQL query intended to retrieve the target construction within a corpus analysis tool (noSketch Engine). The corresponding prompt (*A*) was formulated as follows: *"Write a valid CQL query that could be used to find Croatian constructions of this type in the corpus."*

The GPT-4o model readily proposes a wide range of exploratory strategies and CQL query sketches for identifying repetitive patterns. In the following, part of the results related to the construction type `'X za X'` in Croatian corpora are reported. LLM productively suggests parametric variation across part-of-speech classes, case inflection, and additional morphological or positional constraints (e.g., attempts to enforce identical lemmas across the flanking nouns) (see Table 3). However, many of the automatically generated queries are syntactically imprecise relative to the CQL syntax supported by noSketch Engine and therefore cannot be executed without manual revision. *Query 1* in Table 3 is the first baseline attempt by the model but is syntactically invalid and therefore unusable. Once the model is explicitly instructed that the query must run in noSketch Engine, it produces *Query 2*, which is syntactically correct but still fails to capture the intended semantics of repetition or intensification. *Query 3* represents an autonomous suggestion by the model to constrain the construction to a nominative–instrumental case pair—presumably to increase the likelihood of retrieving syntactically parallel noun sequences. *Query 4* is the first fully executable pattern. However, it returns semantically irrelevant results such as *lijek za upalu* 'medicine for inflammation' or *odjel za poljoprivredu* 'department for agriculture', which do not correspond to the iterative or emphatic use of `X za X`. Since the model does not recognize this mismatch, it proposes *Query 5* as a solution—fixing a single lemma of interest (*rat 'war'*) to narrow the results. While this query is syntactically and semantically valid, it overly restricts the search space.

A different LLM (o3) generates *Query 6* for the same prompt (*B*). This pattern captures the intended semantics (requiring lemma identity) but suffers from invalid syntax and cannot be executed. However, once o3 is given a single example of a valid query via one-shot prompting, it successfully produces *Query 7*, which is both syntactically correct and semantically appropriate.

Overall, while LLMs are useful for exploring strategies and sketching query logic, they lack sufficient grounding in formal CQL syntax specific to noSketch Engine. Expert intervention remains crucial to convert generated ideas into syntactically valid and semantically precise corpus queries. Due to the lack of training material on the CQL language, the LLMs employed here clearly do not possess adequate knowledge of the syntax used in the noSketch Engine implementation. Nevertheless, a positive finding is that few-shot prompt learning and more advanced prompt engineering techniques can yield promising results.

6.2 Data retrieved from corpus searches

To determine how the German repetitive construction `X über X 'X upon X / X over X'` is most typically realized in Croatian, we tested three CQL queries on the CLASSLA-web.hr corpus using different prepositions: *preko* (lit. *over*), *na* (lit. *on*) and *za* (lit. *for*). Each query targeted constructions in which the same noun appears on both sides of the preposition, either in the same or different grammatical case, depending on the structure.

The first query was:

```
1:[tag="N..p.*"] 2:[word="preko"] 3:[tag="N..p.*"] & 1.lemma=3.lemma
```

This query was designed to retrieve instances of the pattern `X preko X ('X over X')`, with the repeated noun in the plural form. However, this structure did not yield examples

No.	CQL Query	Correct Syntax	Correct Semantics	LLM
1	`[upos="NOUN" & number="Sing" & case="Nom"] [lemma="za" & upos="ADP"] [lemma=".*" & upos="NOUN" & number="Sing" & case="Ins" & lemma="\1"]`	No	No	GPT-4o
2	`[tag="N.*" & tag!="Np.*"] [word="za" & tag="S.*"] [tag="N.*" & tag!="Np.*"]`	Yes	No	GPT-4o
3	`[msd="N..s.*n.*"] [word="za" & msd="S.*"] [msd="N..s.*i.*"]`	No	No	GPT-4o
4	`[tag="N.*" & tag!="Np.*"] [word="za" & tag="S.*"] [tag="N.*" & tag!="Np.*"]`	Yes	No	GPT-4o
5	`[lemma="rat" & tag="N.*"] [word="za"] [lemma="rat" & tag="N.*"]`	Yes	Yes/No	GPT-4o
6	`1:[tag="N.*"] [lemma="za"] 2:[tag="N.*" & lemma="1.lemma"]`	No	Yes	o3
7	`1:[tag="N...n*"] 2:[word="za"] 3:[tag="N...i*"] & 1.lemma=3.lemma`	Yes	Yes	o3

Table 3: Evaluation of CQL queries generated by LLMs

corresponding to the expected repetition. Most retrieved examples had different meanings altogether, for instance:

Upoznajte nove ljude preko ljudi koje poznajete. ('Meet new people through [lit. over] people you already know.')

These instances reflect relational uses of *preko* (lit. over), not target reduplication.

The second query tested the pattern `X na X 'X on X'`:

```
1:[tag="N..p.*"] 2:[word="na"] 3:[tag="N..p.*"] & 1.lemma=3.lemma
```

This search retrieved constructions of the form [plural noun] na (lit. on) [same plural noun]. While the results yielded some surface-level matches, they mostly represent literal or spatial constructions rather than idiomatic repetition. Examples include:

usta na usta ('mouth to mouth')
vrata na vrata ('door to door')

Only one attestation matched the intended type of semantic emphasis:

...da ne uzvraćaju uvredama na uvrede... ('...not to respond to insults with insults...')

Finally, the third and most productive query focused on the `X za X-om (lit. X for X) 'X after X'` pattern:

```
1:[tag="N...n*"] 2:[word="za"] 3:[tag="N...i*"] &1.lemma=3.lemma
```

This query searched for constructions in which a noun in the nominative is followed by *za* ('for') and the same noun in the instrumental form (accusative + instrumental construction). The results were far more robust and aligned with the intended meaning of sequential repetition. The five most frequent constructions retrieved from the corpus are shown in Table 4.

Rank	**Construction**	**Frequency (hits)**
1	*dan za danom* 'day after day'	1964
2	*godina za godinom* 'year after year'	123
3	*hit za hitom* 'hit after hit'	112
4	*uspjeh za uspjehom* 'success after success'	99
5	*tjedan za tjednom* 'week after week'	84

Table 4: Most frequent Croatian instances of the `X za X-om 'X after X'` construction based on corpus data

6.3 LLM-classified concordance data

To evaluate the broader productivity and semantic scope of the repetitive construction `X za X` (lit. X for X) in Croatian, a CQL query was first executed on the CLASSLA-web.hr

corpus, retrieving over 5,000 concordance lines. The query targeted same-lemma repetitions across *za* with a nominative singular noun followed by the same lemma in the instrumental singular (e.g., *dan za danom* 'day after day', *afera za aferom* 'affair after affair'). From this large dataset, a random sample of 100 sentences was selected using the *Get a random sample* function in noSketch Engine. For a pilot, a sample of this size provides a sufficiently reliable rough estimate while keeping manual adjudication manageable. As a precaution, we first tested the prompt on this small random batch to confirm output quality before applying it to the full set, a standard pilot-first step in LLM-assisted annotation. The random selection also guards against selection bias, yielding unbiased estimates of the share of idiomatic instances and of the LLM's filtering performance.

An LLM was then employed to distinguish idiomatic usages of the construction from literal or structurally similar, but semantically unrelated occurrences. A carefully engineered prompt guided the classification (see Appendix, prompt *B*). The model was instructed to identify only those instances in which `'X za X'` functions as a repetitive idiomatic construction – expressing rhetorical emphasis, intensification, succession, or accumulation – as opposed to literal pairings or concrete exchanges.

The output was structured in tabular form with three columns: (1) the sentence, (2) a binary label (*valid* or *not valid*), and (3) a short justification. The prompt included positive and negative examples, detailed definitions, and strict classification criteria.

On a random 100-sentence sample, the LLM labeled 79/100 instances of `'X za X'` as idiomatic and 21/100 as non-idiomatic. Expert adjudication judged 99/100 idiomatic and 1/100 literal (a book title *Život za životom* 'Life after life'). In comparison to this gold standard, the LLM achieved an accuracy rate of 80%. All errors were false negatives: the model was conservative in title- or headline-like contexts with minimal cues of iteration, leading it to miss idiomatic uses.

This error profile indicates a conservative bias (no false positives but missed idiomatic uses in headline-like, low-context sentences). Therefore, before scaling to 5,000 instances, we will refine the prompt with in-domain positive examples and route *not valid* cases to brief human review.

In conclusion, this workflow may transfer to other construction types with minor prompt retuning, offering a scalable and adaptable strategy for large-scale phraseological data.

7. Discussion

The pilot study demonstrated both the potential and limitations of combining LLMs and corpus tools for cross-linguistic identification of phraseme constructions. One of the key findings was the differential performance between GPT-4o and the o3 model. While GPT-4o provided semantically plausible suggestions, its output was occasionally influenced by the source language (German), leading to structurally unnatural or less idiomatic Croatian equivalents. In contrast, o3 displayed greater sensitivity to Croatian morphosyntactic constraints, successfully identifying the productive `X za X-om pattern ('X after X')` and explaining its idiomatic function.

However, the study also revealed technical and linguistic limitations of current LLMs. In particular, both models struggled with generating fully valid CQL expressions compat-

ible with the CQL implementation used in noSketch Engine. Errors included incorrect syntax (e.g., embedding variable references inside quotation marks; the incorrect use of positional variable references (e.g., `1:`, `2:`, `3:`)) and an inadequate grasp of Croatian-specific morphological tags under the MULTEXT-East specification. These shortcomings required human correction, reaffirming the necessity of expert oversight in LLM-assisted workflows. Nevertheless, the findings also suggest that even minimal intervention—such as one-shot or few-shot prompt learning—can enable LLMs to generate syntactically valid and semantically relevant CQL queries.

Despite these issues, several results exceeded expectations. For instance, the o3 model provided nuanced comparisons of Croatian variants (e.g., `X na X vs. X za X-om`), including morphosyntactic insights and pragmatic commentary, which enriched the overall interpretation. The model's ability to articulate how prepositions modulate meaning suggests a promising avenue for integrating LLMs into future lexicographic or pedagogical tools.

The corpus findings confirmed that, among the tested patterns, only the `X za X-om 'X after X'` construction occurs with significant frequency in Croatian, fulfilling both the structural and semantic criteria of the German `X über X` prototype. This supports the claim that repetition-based phraseme constructions, while cross-linguistically comparable, are subject to language-specific constraints in productivity, case usage, and prepositional pairing.

Finally, the study illustrated the importance of iterative, layered methodologies that combine human linguistic knowledge with computational assistance. While LLMs can generate plausible hypotheses and pattern abstractions, corpus verification remains essential to validate those outputs against actual usage. The human-in-the-loop framework thus offers a robust strategy for identifying and modeling phraseme constructions in a multilingual setting.

8. Conclusion

This pilot study investigated the potential of integrating AI tools with corpus-based methods for the identification and analysis of phraseme constructions across languages. Focusing on the repetitive construction `X über X` in German, the research tested whether LLMs such as GPT-4o and o3 could generate structurally and semantically plausible Croatian equivalents and formulate accurate CQL expressions for corpus-based validation.

The results demonstrate that while LLMs can generate insightful suggestions and even propose suitable CQL queries, their output often requires expert refinement, particularly in tasks involving formal syntactic precision and language-specific idiomaticity. The o3 model showed higher accuracy than GPT-4o, especially in identifying the Croatian construction `X za X-om 'X after X'` as the most appropriate equivalent of the German model. Corpus analysis using CLASSLA-web.hr further confirmed the productivity and idiomatic status of this Croatian pattern.

By combining AI-generated proposals with empirical data from corpora and expert linguistic oversight, the study outlines a scalable strategy for identifying cross-linguistic phraseme constructions. It has successfully met all three objectives stated at the outset of the study. This hybrid methodology proved effective in balancing semantic abstraction with usage-based validation.

Future research should extend this strategy to additional construction types and language pairs within the PhraConRep framework. It may also explore the integration of the resulting Croatian-German idiom alignments into existing multilingual resources, such as the LIdioms dataset. This LLOD-formatted repository already includes Croatian idioms alongside those from English, German, Italian, Portuguese, and Russian, but currently contains relatively few explicit cross-linguistic links (Filipović Petrović et al., 2024). The alignments produced in this study could enrich the dataset by contributing new connections between Croatian and German idioms.

9. Acknowledgements

This research was supported by the project *Hybrid AI Approaches to Natural Language Processing and Knowledge Generation – HyAI* (uniri-iz-25-215), funded by the European Union – NextGenerationEU.

Software

AbuMandour, W. (2024). Empowering bilingual lexicography with AI: The role of AI-model Triangulation Approach (AMTA) in dictionary design. In A. Inoue, N. Kawamoto & M. Sumiyoshi (eds.) *The 17th International Asian Association for Lexicography Conference*. Tokyo, Japan, pp. 70–83.

Beliga, S. & Filipović Petrović, I. (2024). Large Language Models Supporting Lexicography: Conceptual Organization of Croatian Idioms. In Š. Arhar Holdt & T. Erjavec (eds.) *Proceedings of the Conference on Language Technologies and Digital Humanities*. Ljubljana: Institute of Contemporary History, pp. 23–46.

Beliga, S. & Filipović Petrović, I. (2025). Can AI understand Croatian idioms? Assessing large language models in lexicographic tasks. *Prispevki za novejšo zgodovino*. Accepted for publication.

Chen, L., Dao, H.L. & Do-Hurinville, D.T. (2024). AI empowerment: Where are we in the automation of lexicography? A metaphraseographic study. In A. Inoue, N. Kawamoto & M. Sumiyoshi (eds.) *The 17th International Asian Association for Lexicography Conference*. Tokyo, Japan, pp. 90–98.

Davies, M. (2025). Integrating AI / LLMs into English-Corpora.org. https://www.english-corpora.org/ai-llms/. Accessed: June 15, 2025.

Dobrovol'skij, D. (2018). Sind Idiome Konstruktionen? In K. Steyer (ed.) *Sprachliche Verfestigung. Wortverbindungen, Muster, Phrasem-Konstruktionen*, number 79 in Studien zur deutschen Sprache. Leibniz-Institut für Deutsche Sprache (IDS) [Zweitveröffentlichung], pp. 11 – 23.

Dobrovol'skij, D. (2011). Phraseologie und Konstruktionsgrammatik. In A. Lasch & A. Ziem (eds.) *Konstruktionsgrammatik III: Aktuelle Fragen und Lösungsansätze*. Stauffenburg, pp. 111–130.

Dobrovol'skij, D. (2022). *Deutsche Phrasem-Konstruktion [X hin, X her] in kontrastiver Sicht: eine korpusbasierte Analyse*. Berlin, Boston: De Gruyter, pp. 225–246. URL https://doi.org/10.1515/9783110770209-009.

Filipović-Petrović, I. & Beliga, S. (2024). Lexicographic Treatment of Idioms and Large Language Models: What Will Rise to the Surface? In S. Krek (ed.) *Book of Abstracts of the Workshop Large Language Models and Lexicography*. Ljubljana: Centre for language resources and technologies, University of Ljubljana, pp. 12–16.

Filipović Petrović, I., López Otal, M. & Beliga, S. (2024). Croatian Idioms Integration: 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)*. Torino, Italia: ELRA and ICCL, pp. 4106–4112. URL https://aclanthology.org/2024.lrec-main.366/.

Gantar, A. (2024). Formulisanje riječničkih definicija pomoću umjetne inteligencije na primjeru slovenskih frazeoloških jedinica. In S. Marjanović (ed.) *Moderni riječnici u funkciji prosečnog korisnika: stari problemi, suvremeni pravci i novi izazovi*, volume 1. Beograd: Filološki fakultet Univerziteta u Beogradu, pp. 151–157.

Goldberg, A.E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. Chicago and London: The University of Chicago Press.

Goldberg, A.E. (2019). *Explain Me This: Creativity, Competition and the Partial Productivity of Constructions*. Princeton and Oxford: Princeton University Press.

Hohenhaus, P. (2004). Identical Constituent Compounding – a Corpus-based Study. *Folia Linguistica*, 38(3-4), pp. 297–332. URL https://doi.org/10.1515/flin.2004.38.3-4.297.

Horn, L.R. (2018). *The lexical clone: Pragmatics, prototypes, productivity*. Berlin, Boston: De Gruyter Mouton, pp. 233–264. URL https://doi.org/10.1515/9783110592498-010.

Li, S., Chen, J., Yuan, S., Wu, X., Yang, H., Tao, S. & Xiao, Y. (2024). Translate Meanings, Not Just Words: IdiomKB's Role in Optimizing Idiomatic Translation with Language Models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), pp. 18554–18563. URL https://ojs.aaai.org/index.php/AAAI/article/view/29817.

Ljubešić, N. & Kuzman, T. (2024). CLASSLA-web: Comparable Web Corpora of South Slavic Languages Enriched with Linguistic and Genre Annotation. 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)*. Torino, Italia: ELRA and ICCL, pp. 3271–3282. URL https://aclanthology.org/2024.lrec-main.291/.

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

Mellado Blanco, C. (2022). Phraseology, patterns and Construction Grammar: An introduction. In C. Mellado Blanco (ed.) *Productive Patterns in Phraseology and Construction Grammar: A Multilingual Approach*. Berlin, Boston: De Gruyter, pp. 1–26. URL 10.1515/9783110520569-001.

Mellado Blanco, C. (2023). From idioms to semi-schematic constructions and vice versa: The case of [a un paso de X]. In I. Hennecke & E. Wiesinger (eds.) *Constructions in Spanish*. John Benjamins Publishing Company, pp. 103–128. URL 10.1075/cal.34.05mel.

OpenAI (2024). GPT-4o System Card. URL https://arxiv.org/abs/2410.21276. 2410.21276.

OpenAI (2025). o3-mini System Card. OpenAI technical documentation. Available at https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

Pavlova, A. (2024). Äquivalenz bei Übersetzung von Phrasemkonstruktionen. In A. Gondek, A. Jurasz, M. Kałasznik, P. Staniewski, J. Szczęk & A. Kamińska (eds.) *Interkulturelles und Interdisziplinäres in der Phraseologie und Parömiologie. Bd. II*. Hamburg: Verlag Dr. Kovač, pp. 159–178.

Pavlova, A., Naiditch, L. & Pöppel, L. (2022). Est' kreativnost' i kreativnost'. O granjah i granicah kreativnosti v oblasti frazeologizmov-konstrukcij. *Anzeiger für Slavische Philologie*, 50(1), pp. 181–204.

PhKW (2024). PhKWB Repository: Artikel. https://github.com/PhKW/PhKWB/tree/main/Artikel. Last updated: November 3, 2024; Accessed: April 1, 2025.

Piunno, V. (2022). Coordinated constructional intensifiers: patterns, function and productivity. In C. Mellado Blanco (ed.) *Productive Patterns in Phraseology and Construction Grammar: A Multilingual Approach*. Berlin, Boston: De Gruyter, pp. 133–164.

Shalevska, E., Kostadinovska-Stojchevska, B., Janusheva, V., Janusheva, M., Stojanoska, M. & Talevska, M. (2025). AI IN NONLITERAL LANGUAGE TRANSLATION: TRANSLATING MACEDONIAN PROVERBS AND IDIOMS. *International journal of Education Teacher*, 29, p. 60–66. URL https://www.ijeteacher.com/index.php/ijet/article/view/83.

Sørensen, N.H. & Nimb, S. (2025). The Danish Idiom Dataset: A collection of 1000 Danish idioms and fixed expressions. In H. Einarsson, A. Simonsen & D.S. Nielsen (eds.) *Proceedings of the 1st Workshop on Nordic-Baltic Responsible Evaluation and Alignment of Language Models (NB-REAL 2025)*. Tallinn, Estonia: The University of Tartu Library, pp. 55–63. URL https://aclanthology.org/2025.nbreal-1.5/.

Steyer, K. (2013). *Usuelle Wortverbindungen. Zentrale Muster des Sprachgebrauchs aus korpuslinguistischer Sicht*. Number 65 in Studien zur deutschen Sprache. Tübingen: Narr Francke Attempto Verlag.

Ziem, A. (2018). Construction Grammar meets Phraseology: eine Standortbestimmung. *Linguistik Online*, 90(3). URL https://doi.org/10.13092/lo.90.4316.

Appendix

Prompt A:

- *Explain what these German sentences mean? List of selected sentences: [... 32 original sentences from the PhraConRep repository ...]*
- *Based on the entire set of sentences, identify and describe the repeated phraseological construction that they instantiate.*
- *Provide examples of this construction in Croatian?*
- *Write a valid CQL query that could be used to find Croatian constructions of this type in the corpus.*
- *Adapt the query syntax to be suitable for use in the noSketch Engine tool.*

Prompt B:

- **Persona:** *You are an expert linguist specializing in idioms, phraseology, and cross-linguistic constructions.*
- **Task:** *Analyze the provided list of Croatian phraseme constructions. Your task is to determine whether the pattern X za X-om (lit. X for X) in each sentence functions as a repetitive idiomatic construction or whether it is used literally or non-idiomatically. You are looking for phrases where the repetition of a noun using the pattern X za X-om expresses succession, piling up, or accumulation of the same type of entity. These constructions may express overwhelming quantity, intensity, continuity, or rhetorical emphasis, and they are functionally equivalent to the German construction X über X (Fehler über Fehler, Tage über Tage).*

- **Definition of repetitive construction:** *A repetitive construction is a multi-word expression where a word or a morphological/syntactic pattern is intentionally repeated to create emphasis, intensity, rhythm, or denote totality, iteration, or specific nuanced meaning.*
- **Examples of valid use:**
  - *Bio je to pakleni krug iz kojeg nije bilo izlaza, svađa za svađom – sa ženom, braćom, sestrama, roditeljima, stalni nesporazumi. → repeated conflicts*
  - *Postala je zvijezda dok je još bila tinejdžerica, a reda uspjeh za uspjehom. → repeated successes*
  - *Godina za godinom, porodica se proširivala, dobili su dva sina i navikli na život u novoj sredini. → successive years, marks temporal continuity and development*
- **Examples of not valid use:** *Knjiga Život za životom opetovano prati Ursulu Todd kroz burna zbivanja prošlog stoljeća.*
- **Input Format:** *You will receive 100 sentences (from a corpus), each containing a construction of the form X za X-om.*
- **Output Format (table with 3 columns):**
  1. *Sentence – full original sentence*
  2. *Label – 'Valid' or 'Not valid'*
  3. *Justification – brief explanation (max. 25 words)*
- *Be selective and strict. If the idiomatic usage is not clear, choose "Not valid".*