Injecting Structured Lexicographic Knowledge into LLMs for Non-Literal Expression Disambiguation: A Controlled Study on Croatian

Slobodan Beliga[1,2], Ivana Filipović Petrović[3], Ana Meštrović[1,2]
[1]Faculty of Informatics and Digital Technologies, University of Rijeka, Rijeka, Croatia
[2]Center for Artificial Intelligence and Cybersecurity, University of Rijeka, Rijeka, Croatia
[3]Croatian Academy of Sciences and Arts, Zagreb, Croatia
sbeliga@inf.uniri.hr, ifilipovic@hazu.hr, amestrovic@inf.uniri.hr

Abstract

In potentially idiomatic expressions (PIEs), the same surface form may receive either a literal or an idiomatic interpretation depending on context, making automatic literal–idiomatic disambiguation challenging. This is acute for Croatian, where annotated data and locally runnable generative models are limited. We present a study of Croatian PIE literal–idiomatic disambiguation examining how structured lexicographic knowledge can improve open-weight, decoder-only LLMs without fine-tuning. Using a new expert-annotated concordance dataset – CroPIEs, we compare baseline prompting to inference-time knowledge injection via retrieval-augmented generation (RAG) from a Croatian phraseological dictionary. We isolate the contribution of three knowledge types: definitional knowledge (structured meanings), contextual knowledge as curated prototypical usage examples, and their combination. Results show consistent improvements in macro-F1 for both GaMS-2B-Instruct and GaMS-9B-Instruct models. Definitional knowledge is generally more stable than examples alone, while examples can be effective but less consistent across expressions. The strongest and most reliable gains are obtained when definitions and examples are combined, indicating a synergistic effect between explicit meaning descriptions and contextual cues. Per-class analyses show that injected lexicographic evidence mitigates baseline biases between LITERAL and IDIOMATIC predictions, improving decision balance in a low-resource setting with small data of compact, expert-curated lexicographic evidence injected at inference time.

Keywords: literal–idiomatic disambiguation, PIEs, idioms, structured lexicographic knowledge, knowledge injection, LLM, RAG, GaMS

1. Introduction

Non-literal expressions (NLEs), including many phraseological units (PUs), have long been recognized as a persistent challenge for natural language processing(Sag et al., 2002; Baldwin and Kim, 2010; Constant et al., 2017). This tension is especially visible when the same surface string admits both a literal and a figurative reading, making context-sensitive semantic disambiguation unavoidable. A central task in this area is *literal–idiomatic usage disambiguation*: determining whether a given occurrence of a PU is used compositionally or figuratively.

In corpus settings, many candidates can occur either idiomatically or literally depending on context. Following Haagsma et al. (2020), we adopt the term *potentially idiomatic expressions* (PIEs), defined as expressions that *can* have an idiomatic meaning regardless of whether they actually have that meaning in a given context. In other words, the same surface form may receive either a compositional (literal) reading or a non-compositional (idiomatic/figurative) reading depending on contextual cues. Literal–idiomatic disambiguation is therefore a context-sensitive decision that affects meaning preservation in downstream applications such as machine translation, summarization, and information extraction, where misinterpreting idioms can distort meaning.

This decision remains brittle for large language models (LLMs) in practice. Literal occurrences are frequently over-labeled as idiomatic (Tedeschi et al., 2022), and conversational prompting can induce agreement biases unless the output protocol is tightly constrained (De Luca Fornaciari et al., 2024). Conversely, structured prompts with constrained outputs can yield competitive performance on PIE identification without task-specific fine-tuning (Hashiloni et al., 2025), yet benchmark gains may still rely on spurious cues rather than grounded interpretation (Chakrabarty et al., 2022). For less-resourced languages, dictionaries and phraseological repositories provide compact, vetted semantic knowledge, and definitions and curated examples can be exploited as task guidance (Škvorc and Robnik-Šikonja, 2025).

We therefore study inference-time grounding via retrieval-augmented generation (RAG), which injects external knowledge without additional training. Prior evidence suggests that dictionary-augmented prompting can help, but gains depend on coverage

Proceedings of Learning Non-Literal Expressions with Small Data @ LREC 2026, pages 21–30
11 May 2026.

and ambiguity handling (Perak et al., 2024). More broadly, RAG is often more cost-effective than parameter updates, yet its impact hinges on retrieval quality and context construction, motivating controlled evaluations that isolate the contribution of curated external knowledge (Abonizio et al., 2025; Bhushan et al., 2025).

Against this background, we present a controlled study of Croatian PIE disambiguation, focusing on idioms as a salient PU subclass. Croatian remains comparatively low-resourced: high-quality *local* LLMs are scarce, and available models are typically not trained primarily on Croatian but only adapted to it, which makes inference-time grounding especially attractive. Our primary goal is to quantify, in a *small data* setting, how much *structured lexicographic knowledge from a Croatian phraseological dictionary* helps literal–idiomatic disambiguation: specifically, the contribution of *(i)* definitional knowledge, *(ii)* curated prototypical usage contexts (examples), and *(iii)* their synergy when both are injected via RAG. Using a new expert-annotated Croatian concordance dataset (10 idioms; 1,000 instances), we compare baseline prompting to three knowledge variants (definitions, examples, and their combination) under identical prompting and decoding constraints, and check robustness on two smaller open-weight decoder-only GaMS models of markedly different sizes (2B vs. 9B). Our contributions are threefold:

1. We release a new expert-annotated Croatian concordance dataset for literal–idiomatic usage disambiguation (1,000 instances; 10 PIEs/idioms).
2. We conduct a controlled evaluation of inference-time lexicographic knowledge injection via RAG on two Croatian-adapted open-weight decoder-only LLMs (GaMS-2B-Instruct and GaMS-9B-Instruct), comparing baseline prompting to three knowledge variants (definitions, usage examples, and their combination).
3. We provide empirical evidence that small, manually curated phraseological resources can improve literal–idiomatic disambiguation and mitigate class-level prediction biases in a low-resource setting, with gains supported by statistical testing.

2. Related Work

Automatic idiom detection and usage disambiguation has traditionally been framed as a binary classification task distinguishing literal from idiomatic usage of identical surface forms. Neural approaches such as MICE (Škvorc et al., 2022) demonstrated that contextual embeddings (e.g. ELMo, BERT) encode signals sufficient for detecting idiomaticity, including cases involving unseen expressions. These findings confirmed the importance of contextualized representations, while also underscoring challenges related to limited annotated data and generalization across idioms.

More recently, large language models have been evaluated on idiomatic and figurative language understanding in prompting-based setups. The DICE benchmark (Mi et al., 2025) investigates LLM-based idiom comprehension and explanation, focusing on generative interpretation across model sizes and prompting strategies. Broader figurative reasoning datasets such as FLUTE (Chakrabarty et al., 2022) further explore non-literal interpretation and contextual inference. Together, these studies indicate that pretrained models capture substantial figurative knowledge, although performance remains sensitive to task formulation and prompting design.

Within the Croatian context, recent research has examined the role of LLMs in lexicographic and phraseological tasks. Studies have explored LLM-assisted conceptual organisation of idioms and semantic grouping within the *Online Dictionary of Croatian Idioms* (Beliga and Filipović Petrović, 2024; Filipović Petrović and Beliga, 2025), as well as AI- and corpus-based strategies for identifying phraseme constructions through hybrid human–LLM workflows (Beliga and Filipović Petrović, 2025). These developments highlight the growing integration of AI tools, corpus technologies, and structured lexicographic resources in Croatian phraseological research.

3. Data

This section describes the CroPIEs-1k concordance dataset and the structured lexicographic resource used for RAG.

3.1. Concordance Dataset

Corpus Source. The dataset was derived from the Croatian web corpus CLASSLA-web.hr 2.0 CLARIN.SI (2024), compiled from the national *.hr* domain (2021–2024) and available via CLARIN.SI through the NoSketch Engine concordancer. The corpus covers heterogeneous web genres (e.g. news, blogs, forums) and provides morphosyntactic annotation and advanced querying, enabling precise extraction of PIE candidates. Importantly, all instances in our dataset were selected on the basis of attested usage in authentic corpus data (i.e., they are not synthetically generated).

Selection of Phraseological Units. Ten verb-based phraseological units were selected (Table 2) to ensure: (i) attestation in contemporary Croatian,

(ii) identical surface form in literal and idiomatic usage, (iii) sufficient frequency in CLASSLA-web.hr 2.0 (each >1,000 occurrences), and (iv) a clear semantic contrast between compositional (literal) and non-compositional (idiomatic) readings. The requirement of identical lexical form ensures that the task involves semantic disambiguation within the same surface string. In all selected cases, the same lexical sequence may receive either a literal, compositionally interpretable reading or an idiomatic, phraseological interpretation depending on context. The expressions exhibit syntactic variability, including inflectional variation (tense, aspect, agreement), word-order alternations, insertion of modifiers, and clitic placement, reflecting the rich morphosyntactic structure of Croatian. Such variation is an inherent property of phraseological usage in context and does not represent deviation from a canonical form. Disambiguation therefore requires contextual semantic interpretation across naturally occurring structural variants. All expressions are specified in their canonical form (Table 1), while their corpus attestations reflect authentic grammatical realizations. All selected expressions are well established in contemporary usage and display substantial corpus frequency, ensuring that the dataset reflects productive language patterns.

Concordance Extraction. For each expression, concordance lines were extracted in KWIC format with ±100 characters of left/right context. Given the high frequency of each expression (>1,000 occurrences), we used the NoSketch Engine random sampling function to select 100 instances per expression (10×100=1,000 total). This function generates a representative subset of concordance lines while preserving corpus distribution across sources and genres. The choice of 100 instances per expression ensures sufficient contextual variability for reliable idiom-level evaluation while maintaining uniform sample size across expressions. The resulting dataset, CROPIES-1K, will be released publicly upon acceptance (link in the camera-ready version).

Manual Annotation. All instances were manually annotated by an expert linguist in Croatian phraseology as LITERAL (compositional) or IDIOMATIC (phraseological). While inter-annotator agreement was not measured due to the single-annotator setup, annotation decisions followed consistent criteria based on contextual semantic interpretation. The relatively clear distinction between literal and idiomatic usage in the selected dataset reduces the likelihood of systematic ambiguity. Overall, the task was generally straightforward for an expert annotator.

Table 1 confirms near-balance across classes, reducing confounding effects due to class imbalance in evaluation.

Table 1: Class distribution across PU_{1-10} ($n = 100$ per subset; total $N = 1000$ in CroPIEs-1k dataset.)

PU	1	2	3	4	5	6	7	8	9	10	$\mu \pm \sigma$
IDM	48	51	45	50	49	52	47	50	48	51	**49.1 ± 2.1**
LIT	52	49	55	50	51	48	53	50	52	49	**50.9 ± 2.1**

3.2. Lexicographic Resource for RAG

The external knowledge for RAG was drawn from the *Online Dictionary of Croatian Idioms*[1] Croatian Academy of Sciences and Arts (2023), an open-access, corpus-based born-digital resource developed at the Croatian Academy of Sciences and Arts since 2019. Such structured lexicographic resources are particularly valuable for knowledge injection into LLMs, as they provide expert-authored, validated semantic evidence that complements web-derived pretraining data (often including user-generated sources such as Wikipedia and blogs) and can help mitigate knowledge gaps and reduce ungrounded generations.

The dictionary combines manual lexicographic analysis with corpus-supported procedures. Entries were compiled in Lexonomy and are based on systematic corpus examination, including frequency analysis and collocational evidence. All definitions and examples were manually selected and edited by lexicographers to ensure representativeness and semantic precision. Version 2 (2023) contains 563 entries covering 1,165 idioms.

For our RAG pipeline, we exported the dictionary from Lexonomy and converted it to JSONL. We created three parallel collections[2] matching our experimental conditions: **(1) DEF** (definitions only), **(2) EXS** (examples only), and **(3) DEF+EXS** (definitions+examples). Each JSONL record includes a stable sense identifier and idiom-level metadata to enable deterministic matching and consistent retrieval across conditions.

All ten phraseological units in our concordance dataset are covered by the dictionary. For each instance, the pipeline retrieves the corresponding entry and injects its structured content into the prompt, typically one or two definitions (for polysemy) and around two curated examples (occasionally one or three for variant forms). In this study, the dictionary thus provides a small-scale but high-quality expert resource whose structured semantic information is injected to support disambiguation in a low-data setting.

[1] https://lexonomy.elex.is/frazeoloskirjecnikhr

[2] JSONL collections used for retrieval: GITHUB

Expression (Croatian)	Literal Gloss	Idiomatic Meaning (English)
bacati mrvice	to throw crumbs	to offer small concessions deliberately in order to appease someone
okrenuti leđa	to turn one's back	to abandon or withdraw support
isplivati na površinu	to float to the surface	to become visible or publicly known
dati crveni karton	to give a red card	to remove someone from a political or institutional position
graditi mostove	to build bridges	to promote cooperation or reconciliation
biti u sjeni	to be in the shadow	to remain overshadowed or unnoticed
naletjeti na minu	to run into a mine	to encounter an unexpected hidden problem
biti u komi	to be in a coma	to be in a state of lethargy or inactivity
stati na noge	to stand on one's feet	to recover or regain stability
znati koliko je sati	to know what time it is	to know what is going on; to be aware of the situation

Table 2: Phraseological units included in the dataset with literal glosses and idiomatic meanings.

4. Experimental Setup

We evaluate the impact of injecting structured lexicographic knowledge via RAG on Croatian *literal–idiomatic disambiguation*. We compare a baseline condition without retrieval (prompting only) to three RAG variants, using a controlled, paired evaluation design across model sizes.

4.1. Task Formulation and Evaluation

We formulate the task as binary sentence-level *literal–idiomatic disambiguation*. Given a concordance sentence containing a predefined target idiomatic expression (possibly in morphologically or syntactically varied form), the model predicts whether the expression is used LITERAL (compositional) or IDIOMATIC (non-compositional) in that sentence. For each instance x_i, the model outputs $y_i \in \{\text{LITERAL}, \text{IDIOMATIC}\}$; gold labels were assigned manually at the sentence level. Evaluation is strictly categorical (no graded judgments).

Evaluation. We report macro-averaged F1 (macro-F1). Scores are computed (i) globally over all instances and (ii) at the idiom level, where macro-F1 is computed separately for each idiom subset (100 instances). Idiom-level scores are used for paired comparisons across conditions; we report their mean$\pm$SD across the ten idioms. To analyze class-level effects, we also report precision (P), recall (R), and F1 separately for LITERAL and IDIOMATIC (idiom-averaged). For interpretability, we also report Δ macro-F1 relative to the baseline (RAG–baseline) in tables and figures. Statistical significance between baseline and RAG variants is assessed with the Wilcoxon signed-rank test on idiom-level macro-F1 ($n = 10$ idioms), appropriate for the paired and small-sample design. All conditions are evaluated on the same fixed set of instances. A safeguard mechanism was implemented to handle structurally invalid outputs, but no such cases occurred in the reported experiments.

4.2. Models

In this study, we evaluate two open-weight multilingual LLMs from the GaMS (Generative Model for Slovene) family: *GaMS-2B-Instruct*[3] and *GaMS-9B-Instruct*[4]. Both are decoder-only Transformer models based on the Gemma 2 architecture and were continually pretrained and subsequently instruction-tuned for South Slavic languages (Vreš et al., 2024; Vajda et al., 2025). We focus on small, locally runnable open-weight models that can be executed without external APIs on modest GPU hardware, enabling controlled and reproducible experimentation.

The variants differ primarily in scale: GaMS-2B-Instruct has 26 layers (hidden size 2304; 8 attention heads), while GaMS-9B-Instruct has 42 layers (hidden size 3584; 16 attention heads). Both support a maximum context length of 8192 tokens.

GaMS models are not Croatian-specific. While large proprietary LLMs can perform well on Croatian, Croatian remains comparatively low-resourced in terms of openly available, locally runnable instruction-tuned decoder-only LLMs with substantial Croatian coverage. Although primarily developed for Slovene, GaMS was continually pretrained on multilingual corpora that include Croatian (alongside Slovene, Serbian, and Bosnian), making it a practical open-weight choice for controlled experiments on Croatian literal–idiomatic disambiguation.

Comparing 2B and 9B within the same model family keeps the architecture and training paradigm constant while varying capacity, enabling a controlled analysis of how structured lexicographic knowledge injection interacts with model size.

[3] https://huggingface.co/cjvt/GaMS-2B-Instruct

[4] https://huggingface.co/cjvt/GaMS-9B-Instruct

4.3. RAG Configuration

Knowledge injection was implemented via RAG pipeline over a structured lexicographic resource stored in JSONL format. Each record contains a sense identifier, idiom-level metadata, and a textual payload consisting of either (i) a definition (DEF), (ii) curated usage examples (EXS), or (iii) their combination (DEF+EXS), depending on the experimental condition.

Indexing phase. All JSONL records were embedded with `sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`[5] (Reimers and Gurevych, 2019; Wang et al., 2020). The model produces 384-dimensional vectors, which we L2-normalized before indexing. Similarity search uses cosine similarity, implemented as inner-product search over normalized vectors in a FAISS `IndexFlatIP` index.

Retrieval phase. Retrieval is two-stage. We first attempt a deterministic hard match using idiom metadata (exact match to the target idiom). If no direct match is found, we fall back to dense retrieval over the FAISS index. Across all RAG conditions, retrieval parameters are fixed to top-k=1 with a maximum injected context length of 2000 characters.

Injection phase. The retrieved content is inserted into the prompt in a clearly demarcated reference section and treated as the sole external evidence used for the decision. The model is instructed to rely only on the reference text and the concordance sentence (baseline prompting omits the reference section). The retrieval pipeline and all RAG hyperparameters are identical for GaMS-2B-Instruct and GaMS-9B-Instruct, ensuring that differences across conditions stem from the type of injected lexicographic knowledge rather than retrieval configuration. The overall RAG pipeline is summarized in Figure 1.

4.4. Prompting Strategy

We use a single shared prompt template across the baseline and all RAG conditions to ensure strict comparability; the only differences across conditions are (i) whether a reference block is present (baseline: none; RAG: injected content) and (ii) a small condition-specific clarification of what constitutes semantic alignment (definitions vs. examples vs. both). The overall protocol is a structured, conservative two-step decision procedure.

First, the model assigns an alignment score $MAP\_SCORE \in \{0, 1, 2\}$ that reflects how clearly the usage in the concordance matches the available evidence (baseline: sentence context only; RAG: sentence + injected reference).

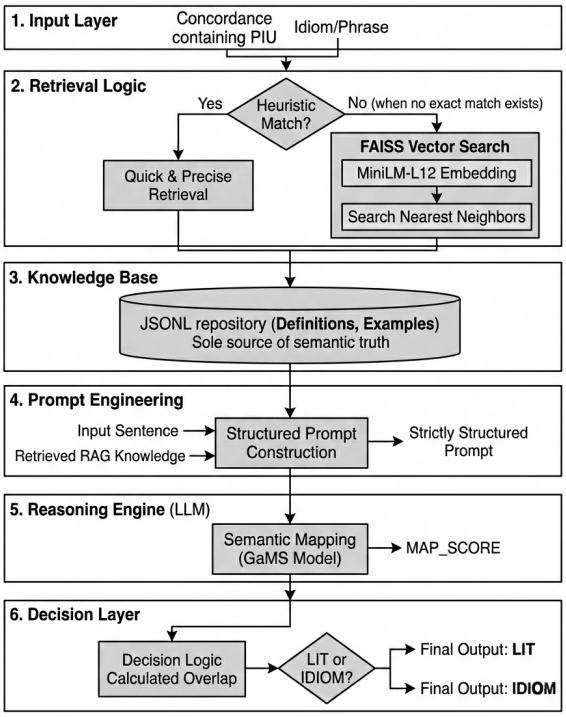

Figure 1: RAG protocol for Croatian literal–idiomatic disambiguation: deterministic metadata match with FAISS fallback, followed by prompt injection of structured lexicographic evidence.

Second, the final label is derived deterministically: only MAP_SCORE=2 yields IDIOMATIC, while $MAP\_SCORE \in \{0, 1\}$ defaults to LITERAL. This mapping enforces conservative grounding: the idiomatic label is allowed only when the model can confidently justify semantic correspondence.

To reduce format variability, all prompts require a fixed three-line output (decision, MAP_SCORE, and a short mapping/justification). We avoided open-ended expert-style prompts (e.g. requesting a linguistic analysis), which increased verbosity and structural variance across model sizes. Larger models also tended to produce more verbose outputs, reinforcing the need for a strictly structured format. The resulting structured prompting setup minimizes confounding effects due to prompt sensitivity and isolates the contribution of injected knowledge. A baseline prompt skeleton is shown below; full Croatian templates for the RAG variants are available online.[6]

```
PROMPT SKELETON
TASK: Decide LITERAL vs IDIOMATIC
for <PIE> in <SENTENCE>.
STEP 1: MAP_SCORE in {0,1,2}
        (semantic alignment)
   - 0: no evidence
   - 1: weak/uncertain evidence
```

[5] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

[6] https://github.com/sbeliga/CroPIEs-1k

```
   - 2: clear evidence
   RESTRICTIONS:
     - sentence(+reference) only;
     - no external knowledge;
     - unsure => MAP_SCORE<2.
STEP 2: If MAP_SCORE=2 -> IDIOMATIC;
        Else -> LITERAL.
[REFERENCE BLOCK - RAG only]
<< injected  Def/Exs/Def+Exs >>
OUTPUT (3 lines):
DECISION; MAP_SCORE; JUSTIFICATION
```

4.5. Inference Configuration

All experiments use controlled decoding to attribute differences to knowledge injection rather than generation variability. We apply greedy decoding (`do_sample=False`, `num_beams=1`) with `repetition_penalty=1.0`, `use_cache=False`, and `max_new_tokens=60`, which is sufficient for the required short, structured outputs. We run mixed-precision inference (FP16) without quantization; a fixed random seed (1234) and deterministic execution are used where supported.

To avoid truncation effects, inputs are tokenized without truncation and a 300-token safety margin is reserved within the context window (fail-fast if exceeded), ensuring the full concordance context is preserved. When available, prompts are rendered with the tokenizer chat template (`apply_chat_template` with `add_generation_prompt=True`) for consistent formatting across models and conditions.

5. Experiments and Results

Table 3 summarizes PIE-level macro-F1 (mean±SD over 10 Croatian PIEs) for GaMS-2B-Ins. and GaMS-9B-Ins. under the baseline and three RAG knowledge-injection variants. Without external knowledge, both models perform modestly, with the larger model outperforming the smaller one (0.4377 vs. 0.3357). RAG improves performance for both models, with the strongest gains consistently obtained by **Def+Exs** (GaMS-2B-Ins.: 0.4584, $\Delta = +0.1227$; GaMS-9B-Ins.: 0.6211, $\Delta = +0.1834$). Definitional knowledge (DEF) yields larger average gains than example-only injection (Exs), suggesting that explicit semantic descriptions provide more stable disambiguation cues than contextual similarity alone. Overall, these results show that a small, manually curated phraseological resource can substantially improve Croatian literal–idiomatic disambiguation in a low-resource setting. Gains are observed even for the smaller model, suggesting that structured lexicographic evidence can partially compensate for limited internal representations in low-resource conditions. The relatively large SD values further indicate substantial heterogeneity across PIEs, motivating an idiom-level analysis.

To test whether gains generalize across PIEs, we use a one-sided exact Wilcoxon signed-rank test on idiom-level differences ($\Delta = F1_{\text{RAG}} - F1_{\text{Baseline}}$, $n = 10$ PIEs) to assess whether the median improvements (Δ) exceeds zero. Holm–Bonferroni correction is applied for multiple comparisons (Table 4). For GaMS-2B-Ins., only **Def+Exs** shows a statistically reliable improvement over the baseline ($p_{\text{Holm}} = 0.0391$, $r = 0.725$). For GaMS-9B-Ins., both DEF ($p_{\text{Holm}} = 0.0146$, $r = 0.822$) and **Def+Exs** ($p_{\text{Holm}} = 0.0059$, $r = 0.886$) are significant, while Exs is not. The effect sizes are large, indicating practically meaningful improvements in addition to statistical reliability.

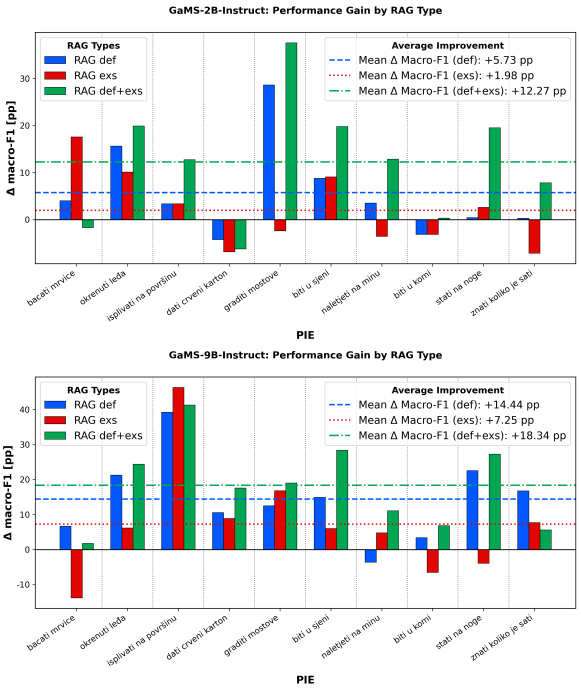

Figure 2: Idiom-level change in macro-F1 (RAG–baseline; percentage points) for 10 Croatian PIEs. Bars show Δ macro-F1 for three knowledge variants (DEF, Exs, DEF+Exs); horizontal lines indicate the mean Δ across PIEs for each variant (within each model).

Figure 2 breaks down these effects by PIE and contrasts the three RAG variants (bars show Δ in percentage points for interpretability; dashed lines denote mean Δ per variant within each model). Improvements are not uniform across expressions, but **Def+Exs** yields the most stable and frequently positive pattern across PIEs, suggesting that combining definitional and contextual evidence provides complementary signals for disambiguation. Notably, for GaMS-9B-Ins. the **Def+Exs** variant does not yield negative changes for any PIE, and over-

Setting	GaMS-2B-Ins.		GaMS-9B-Ins.	
	Macro-F1 (M ± SD)	Δ	Macro-F1 (M ± SD)	Δ
Baseline	0.3357 ± 0.0348	–	0.4377 ± 0.0746	–
RAG (Def)	0.3930 ± 0.1064	+0.0573	0.5821 ± 0.1201	+0.1444
RAG (Exs)	0.3555 ± 0.0530	+0.0198	0.5102 ± 0.1451	+0.0725
RAG (Def+Exs)	**0.4584 ± 0.1424**	**+0.1227**	**0.6211 ± 0.1278**	**+0.1834**

Table 3: PIE-level macro-F1 (mean ± sample SD over 10 Croatian PIEs) for GaMS-2B-Ins. and GaMS-9B-Ins. across the baseline and RAG variants. Δ denotes the mean absolute difference from the baseline, computed on the [0,1] scale.

Model	Config	p_{Holm}	r	Sig.
GaMS-2B-Ins.	Def	0.1260	0.564	–
GaMS-2B-Ins.	Exs	0.3125	0.177	–
GaMS-2B-Ins.	Def+Exs	0.0391	0.725	*
GaMS-9B-Ins.	Def	0.0146	0.822	*
GaMS-9B-Ins.	Exs	0.1934	0.435	–
GaMS-9B-Ins.	Def+Exs	0.0059	0.886	*

Table 4: Wilcoxon signed-rank tests over idiom-level macro-F1 scores ($n = 10$), comparing RAG configurations against the baseline. Holm–Bonferroni correction was applied. Effect size $r = Z/\sqrt{n}$. * indicates $p_{\text{Holm}} < 0.05$.

No.	PIE	Δ 2B	Δ 9B
1	*bacati mrvice*	-0.0168	0.0179
2	*biti u komi*	0.0034	0.0686
3	*biti u sjeni*	0.1980	0.2840
4	*dati crveni karton*	-0.0623	0.1759
5	*graditi mostove*	0.3757	0.1904
6	*isplivati na površinu*	0.1275	**0.4129**
7	*naletjeti na minu*	0.1285	0.1110
8	*okrenuti leđa*	**0.1992**	0.2442
9	*stati na noge*	0.1954	0.2726
10	*znati koliko je sati*	0.0784	0.0561

Table 5: Per-PIE macro-F1 improvement (Δ) of RAG (DEF+EXS) over the baseline for GaMS-2B-Ins. and GaMS-9B-Ins. (computed on the [0,1] scale).

all the 9B model exhibits fewer degradations than 2B across variants. In contrast, Exs is the least stable variant and occasionally produces negative changes, consistent with the intuition that example-only grounding can introduce noise when semantic alignment is weak. The largest gains tend to occur for more semantically opaque PIEs (e.g., *graditi mostove*, *isplivati na površinu*), supporting the usefulness of structured lexicographic knowledge for non-literal interpretation in a low-resource setting.

Table 5 reports per-PIE macro-F1 improvements for the best-performing configuration (DEF+EXS) relative to the baseline. Gains are not uniform across expressions. For GaMS-2B-Ins., DEF+EXS improves performance on 8/10 PIEs, with small drops limited to two cases (*bacati mrvice* and *dati crveni karton*); the largest gains are observed for *graditi mostove* and *okrenuti leđa*. GaMS-9B-Ins. shows more consistent behavior, improving on all 10/10 PIEs and achieving particularly strong gains for *isplivati na površinu* and *biti u sjeni*. Overall, these per-PIE results corroborate the macro-level trends in Table 3 and the distributional patterns in Figure 2, indicating broadly distributed gains rather than isolated outliers.

Table 6 reports idiom-averaged per-class performance, revealing pronounced prediction biases in the baseline condition. GaMS-2B-Ins. exhibits a strong LITERAL bias, achieving near-perfect LITERAL recall (0.9904) while almost completely failing to detect IDIOMATIC usage (R=0.0180, F1=0.0336). In contrast, GaMS-9B-Ins. shows the opposite tendency, strongly favoring IDIOMATIC predictions (R=0.9863) at the expense of LITERAL detection (R=0.1007). These opposing biases highlight that model scale alone does not guarantee balanced literal–idiomatic decisions.

Lexicographic knowledge injection reshapes these class-level behaviors. For GaMS-2B-Ins., all RAG configurations substantially increase IDIOMATIC recall (up to 0.9647 in EXS), confirming that injected evidence helps the smaller model recognize non-literal meaning. However, EXS overgeneralizes toward IDIOMATIC predictions, collapsing LITERAL recall to 0.0268. The combined DEF+EXS variant yields a more balanced trade-off, improving IDIOMATIC F1 (0.5507) while retaining moderate LITERAL performance.

For GaMS-9B-Ins., RAG primarily improves LITERAL recall (from 0.1007 to 0.6011 in DEF+EXS) while maintaining competitive IDIOMATIC performance. Across models, definitional grounding (DEF and DEF+EXS) has a stabilizing effect, whereas example-only injection (EXS) can shift the decision boundary and induce class overprediction in a model-dependent manner. Overall, these results show that structured lexicographic resources improve not only macro-F1 but also the balance between LITERAL and IDIOMATIC predictions by mitigating baseline bias. This is consistent with the conservative MAP_SCORE protocol, where stronger evidence in the reference block can systematically shift the model toward (or away from) the idiomatic decision.

Model	Config	Literal			Idiomatic		
		P	R	F_1	P	R	F_1
GaMS-2B-Ins.	Baseline	0.4750	0.9904	0.6377	0.3000	0.0180	0.0336
	RAG (Def)	0.3046	0.2231	0.2147	0.5024	0.7567	0.5713
	RAG (Exs)	0.2650	0.0268	0.0385	**0.5243**	**0.9647**	**0.6724**
	RAG (Def+Exs)	0.5053	0.3829	0.3660	0.4953	0.6728	0.5507
GaMS-9B-Ins.	Baseline	0.8079	0.1007	0.1741	0.5492	0.9863	0.7013
	RAG (Def)	0.6039	0.5486	0.5208	0.6715	0.6957	0.6434
	RAG (Exs)	**0.5228**	**0.8442**	**0.6337**	0.7126	0.3029	0.3867
	RAG (Def+Exs)	0.6401	0.6011	0.5760	0.6957	0.7026	0.6661

Table 6: Mean per-class performance across idioms. Precision (P), Recall (R), and F1 scores are computed separately for the Literal and Idiomatic classes and averaged over the 10 target idioms for each model and configuration.

6. Discussion

Our results show that inference-time injection of structured lexicographic knowledge can substantially improve Croatian PIE literal–idiomatic disambiguation for locally runnable, open-weight decoder-only LLMs without fine-tuning. Across both GaMS models, RAG yields macro-level gains, with DEF+EXS producing the most reliable improvements and the strongest statistical evidence. This pattern supports the view that definitions and curated usage examples provide complementary signals: definitions act as stable semantic anchors, while examples contribute prototypical contextual cues for matching concordance usage.

At the same time, gains are not uniform across expressions. PIE-level analyses indicate that some expressions benefit strongly, whereas a small subset shows marginal improvements or occasional degradations, highlighting sensitivity to contextual fit and the structure of retrieved evidence. Importantly, the per-class breakdown shows that lexicographic grounding does not merely increase aggregate scores: it can reshape decision behavior by mitigating strong baseline prediction biases (literal-biased GaMS-2B vs. idiomatic-biased GaMS-9B). Example-only grounding is also less stable and can induce class overprediction in a model-dependent way, reinforcing the stabilizing role of definitional evidence.

A limitation of the current setup is that the target PIE can appear in morphologically and syntactically varied realizations within concordances (e.g., inflectional changes, word-order variation, or the insertion of additional words (e.g., modifiers or clitics) within the expression, and occasionally multiple times in the same instance. While our evaluation assumes the target expression is present, further error analysis should disentangle potential failures in (i) identifying the intended target span under such variation and (ii) performing literal–idiomatic disambiguation once the target is identified. Additional confounds arise when semantically related expressions occur nearby: for example, our study targets *dati crveni karton* (*to give a red card*), but some concordances also contain *dati žuti karton* (*to give a yellow card*), which may prime an idiomatic interpretation even though it is not the target of classification. Moreover, a small portion of concordances contains dialectal or informal Croatian; since GaMS is not explicitly trained or optimized for Croatian dialectal varieties or slang (and such data are likely underrepresented), these inputs may reduce both retrieval fit and model comprehension. Finally, structured outputs with brief justification appear more stable than single-label answers, motivating deeper analysis of how explanation requirements interact with grounding and decision reliability.

Despite these limitations, the results consistently indicate that small, expert-curated lexicographic resources provide effective *small data* grounding for non-literal interpretation in low-resource settings. In particular, several PIEs benefit robustly from injected evidence across models and knowledge variants, whereas a few remain challenging even under DEF+EXS, suggesting that targeted analysis of dictionary senses, example selection, and contextual ambiguity could further improve grounding quality.

Overall, these findings align with a small-data, neurosymbolic perspective: compact, expert-curated lexicographic resources can compensate for representational gaps in locally adapted LLMs and provide controlled semantic evidence for non-literal interpretation in low-resource settings. The present findings are derived from a controlled dataset with a limited number of phraseological units. Future work will extend the evaluation to broader PIE inventories and retrieval strategies, and to a larger and more diverse set of idioms, allowing us to evaluate the approach under more heterogeneous conditions.

7. Conclusion

This paper presented a controlled study of Croatian PIE literal–idiomatic disambiguation, evaluat-

ing inference-time injection of structured lexicographic knowledge via RAG on locally runnable, open-weight decoder-only LLMs without fine-tuning. The study isolates the contribution of three knowledge variants: definitions (DEF), curated usage examples (EXS), and their combination (DEF+EXS), under a unified prompting and decoding protocol across two model sizes.

Across models and expressions, lexicographic grounding improves macro-F1, with DEF+EXS yielding the most reliable gains and the strongest statistical evidence. Definitional evidence is more stable than examples alone, while example-only grounding can be less consistent and may shift class-level behavior in a model-dependent way. Beyond aggregate scores, injected knowledge mitigates strong baseline prediction biases and improves the balance between LITERAL and IDIOMATIC decisions.

For reproducibility, we release the CroPIEs-1k expert-annotated concordance dataset and the prompt templates used in this study. Overall, the findings highlight that *small data* in the form of compact, expert-curated structured lexicographic knowledge can provide effective grounding for non-literal language understanding in low-resource settings without fine-tuning.

8. 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. The views and opinions expressed are solely those of the author and do not necessarily reflect the official stance of the European Union or the European Commission. Neither the European Union nor the European Commission can be held accountable for them.

9. Bibliographical References

Hugo Abonizio, Thales Almeida, Roberto Lotufo, and Rodrigo Nogueira. 2025. Comparing knowledge injection methods for llms in a low-resource regime. In *Anais do XXII Encontro Nacional de Inteligência Artificial e Computacional*, pages 819–830, Porto Alegre, RS, Brasil. SBC.

Timothy Baldwin and Su Nam Kim. 2010. Multiword expressions. In Nitin Indurkhya and Fred J. Damerau, editors, *Handbook of Natural Language Processing*, 2 edition, pages 267–292. CRC Press, Boca Raton.

Slobodan Beliga and Ivana Filipović Petrović. 2024. Large language models supporting lexicography: Conceptual organization of croatian idioms. In *Proceedings of the Conference on Language Technologies and Digital Humanities*, pages 23–46, Ljubljana. Institute of Contemporary History.

Slobodan Beliga and Ivana Filipović Petrović. 2025. Ai- and corpus-based strategies for identifying phraseme constructions: A pilot study on croatian repetitive constructions. In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography*, pages 95–115, Brno. Lexical Computing CZ s.r.o.

Kushagra Bhushan, Yatin Nandwani, Dinesh Khandelwal, Sonam Gupta, Gaurav Pandey, Dinesh Raghu, and Sachindra Joshi. 2025. Systematic knowledge injection into large language models via diverse augmentation for domain-specific RAG. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5937–5958, Albuquerque, New Mexico. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mathieu Constant, Gülşen Eryiğit, Johanna Monti, Lonneke van der Plas, Carlos Ramisch, Michael Rosner, and Amalia Todirascu. 2017. Survey: Multiword expression processing: A Survey. *Computational Linguistics*, 43(4):837–892.

Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. A hard nut to crack: Idiom detection with conversational large language models. In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.

Ivana Filipović Petrović and Slobodan Beliga. 2025. Can ai understand croatian idioms? assessing large language models in lexicographic tasks. *Prispevki za novejšo zgodovino*, 65(3):218–242.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 279–287, Marseille, France. European Language Resources Association.

Kai Golan Hashiloni, Ofri Hefetz, and Kfir Bar. 2025. Easy as PIE? identifying multi-word expressions with LLMs. In *Proceedings of the*

2025 Conference on Empirical Methods in Natural Language Processing, pages 23771–23790, Suzhou, China. Association for Computational Linguistics.

Maggie Mi, Aline Villavicencio, and Nafise Sadat Moosavi. 2025. Rolling the DICE on idiomaticity: How LLMs fail to grasp context. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7314–7332, Vienna, Austria. Association for Computational Linguistics.

Benedikt Perak, Slobodan Beliga, and Ana Meštrović. 2024. Incorporating dialect understanding into LLM using RAG and prompt engineering techniques for causal commonsense reasoning. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties, and Dialects (VarDial 2024)*, pages 220–229, Mexico City, Mexico. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. ACL.

Ivan A. Sag, Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2002)*, pages 1–15.

Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. ID10M: Idiom identification in 10 languages. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.

Dario Vajda, Domen Vreš, and Marko Robnik Šikonja. 2025. Improving LLMs for machine translation using synthetic preference data. In *Proceedings of the 2nd LUHME Workshop*, pages 67–73, Bologna, Italy. UP - Universidade do Porto (https://doi.org/10.21747/978-989-9193-73-4/lan2), LIACC - Laboratório de Inteligência Artificial e Ciência de Computadores da Universidade do Porto, CLUP - Centro de Linguística da Universidade do Porto, UEF - The University of Eastern Finland and UAH - Universidad de Alcalá.

Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik-Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters. In *Proceedings of the Conference on Language Technologies and Digital Humanities (JTDH 2024)*, pages 485–511. Institute of Contemporary History, Ljubljana, Slovenia.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Tadej Škvorc, Polona Gantar, and Marko Robnik-Šikonja. 2022. Mice: Mining idioms with contextual embeddings. *Knowledge-Based Systems*, 235:107606.

Tadej Škvorc and Marko Robnik-Šikonja. 2025. Solving word-sense disambiguation and word-sense induction with dictionary examples.

10. Language Resource References

CLARIN.SI. 2024. *CLASSLA-web.hr 2.0*. Croatian web corpus, accessible via NoSketch Engine.

Croatian Academy of Sciences and Arts. 2023. *Online Dictionary of Croatian Idioms (Frazeološki rječnik hrvatskoga jezika), Version 2*. Open-access digital phraseological dictionary.