A Parallel Cross-Lingual Benchmark for Multimodal Idiomaticity Understanding

Dilara Torunoğlu-Selamet[1] **Doğukan Arslan**[1] **Rodrigo Wilkens**[2] **Wei He**[2]

Thomas Pickard[3] **Adriana Silvina Pagano**[4] **Aline Villavicencio**[2,3] **Gülşen Eryiğit**[1]

[1]Istanbul Technical University [2]University of Exeter
[3]University of Sheffield [4]Federal University of Minas Gerais

Relevant UniDive working groups: WG1, WG3, WG4

1 Introduction

Idioms are non-compositional multiword expressions (MWEs) that show some difficulties for both humans and natural language processing (NLP) systems because their meaning is not derivable from their parts. They exhibit shared conceptualizations and cultural knowledge, highlighting a spectrum of transferability which some are shared across languages (e.g., "bad apple" in English and "çürük elma" in Turkish, lit. "rotten apple").

Traditional benchmarks often frame idiom comprehension as a classification task, which does not fully reflect a model's true grasp of idiomatic meaning (Boisson et al., 2023; He et al., 2025). While large language models (LLMs) often struggle with potentially idiomatic expressions (PIEs), especially in low-resource or multilingual scenarios (Arslan et al., 2025), most existing datasets are not constructed in a parallel fashion, preventing direct cross-linguistic comparisons of how idiomatic meaning is preserved or altered across languages.

This work presents XMPIE (Cross-lingual and Multimodal Potentially Idiomatic Expressions), which is a parallel benchmark that covers 34 languages, expands upon the first edition of the AdMIRe shared task framework (Pickard et al., 2025). The dataset includes 3,054 PIEs and 7,040 images depicting a spectrum from figurative to literal meanings, with distractors. It enables a systematic evaluation of whether an LLM's understanding of an idiom in one language or modality can be transferred to another.

2 Related Work

While early resources focused on English compounds (Cook et al., 2008), over time, these were extended to include multilingual entries like PARSEME (Savary et al., 2015) and SemEval-2022 Task 2 (Tayyar Madabushi et al., 2022). Recent datasets have moved beyond simple identification to capture context-dependent aspects of idiomaticity. Examples include DICE (Mi et al., 2025), designed to test how well models use context for disambiguation, and datasets containing minimal pairs and human judgments at both the type and token levels (He et al., 2025). Despite the increase in language coverage, most existing datasets are not constructed in a parallel fashion. This lack of parallelism prevents direct cross-linguistic comparisons and limits research on how idiomatic meaning is preserved or altered during cross-lingual transfer.

There are also researches that explore the interplay between text and images, such as IRFL dataset which established that vision-language models significantly underperform compared to humans on figurative tasks (Yosef et al., 2023). Recent additions like V-FLUTE (Saakyan et al., 2025) require models to provide textual justifications for their visual decisions, while datasets like MChIRC (Wang et al., 2025) focus specifically on Chinese idiom comprehension. Yet, a massively cross-lingually aligned resource for evaluating idiomaticity in both language production and multimodal settings remains missing, which the XMPIE dataset aims to address.

3 Annotation Methodology

The project involved 78 native or highly fluent language experts recruited via the UniDive network. Experts participated in online workshops and were provided with written guidelines and ongoing consultation. Discord was used as the central hub for collaboration, prompt sharing, and creation of image data.

For every seed English PIE selected from first edition of the AdMIRe shared task, annotators provided five textual components: (1) a word-by-word translation into the target language, (2) transliterated version of the literal translation (if applicable), (3) the corresponding idiom in the target language, (4) a literal translation of the target idiom back into English, (5) a transliterated version of the idiomatic form (if applicable).

Then, using Midjourney, annotators generated five images for each PIE: two images for figurative and literal senses, two images weakly related to figurative and literal senses, and an entirely unrelated image. Annotators were instructed to use concrete, visually grounded descriptions rather than abstract terms. The process included "good" and "bad" prompt examples to ensure consistency across the 34 language variants.

4 The XMPIE Dataset

The dataset encompasses 34 language variants and includes over 3,054 expressions derived from English seed PIEs. A total of 7,040 images were generated for these expressions. To maintain quality while ensuring diversity, the languages are categorized into two versions: core dataset consists of 21 language variants (including Azerbaijani, Bulgarian, Chinese, Georgian, Greek, Igbo, Italian, Kazakh, Lithuanian, Norwegian, Portuguese (Brazilian and European), Russian, Serbian, Slovak, Slovenian, Spanish (Ecuadorian and European), Turkish, Ukrainian, and Uzbek) that meet all strict quality thresholds and contain all required modalities; extended dataset includes 13 language variants (Aromanian, Catalan, Danish, Farsi, Hebrew, Hungarian, Indonesian, Javanese, Latvian, Luxembourgish, Macedonian, Swahili, and Urdu) where data collection is ongoing or certain elements like context sentences may still be missing.

In dataset, we identified four primary types of idiomatic alignment between the English PIEs and target languages: (1) 1-1 alignment which corresponds to the target language has an identical or near-identical idiomatic counterpart (e.g., "bad apple" becoming "çürük elma" in Turkish, lit. "rotten apple" in English); (2) 1-0 alignment which corresponds to the target language uses a descriptive (non-idiomatic) expression or a single lexical item rather than a multiword expression (e.g., "chocolate teapot" described as "absolutely useless thing" in Georgian); (3) 1-N alignment which is a single source idiom corresponds to multiple target idioms, often reflecting variation or ambiguity; (4) N-1 alignment which corresponds to multiple English idioms sharing the same equivalent in the target language (e.g., "big fish" and "big cheese" both being rendered as "peix gros" in Catalan).

5 Multilingual and Multimodal Idiomatic Representations

We evaluate system's ability to rank five candidate images for a specific PIE. Each test item includes five standardized image slots: one idiomatic, one literal, two semantically related "weak" variants, and one random distractor. This initial evaluation is "PIE-only," meaning no additional context sentences are provided to help the model disambiguate between literal and figurative meanings.

We utilized EVA-CLIP-18B, an 18-billion parameter vision-language model, as a zero-shot retrieval baseline. The model operates in a joint image-text embedding space, where the PIE string is compared against the images using cosine similarity. The evaluation was conducted across five diverse languages: English (EN), Brazilian Portuguese (BP), Ecuadorian Spanish (ES), Chinese (CN), and Turkish (TR).

In almost all languages (except Spanish), the model showed a massive preference for literal interpretations over idiomatic ones. For example, in English, the literal image was correctly ranked first 90% of the time, while the idiomatic one was only first in 6% of cases. The high NDCG@5 scores across all languages (all > 0.87) indicate that the model generally places relevant images near the top of the list, even if it struggles to distinguish the "exact" idiomatic sense. The very low "Idiomatic Top-2" scores suggest that fine-grained disambiguation remains a significant challenge for current vision-language models without external contextual clues.

6 Conclusion

In conclusion, XMPIE provides a necessary parallel resource for examining how idiomatic concepts are realized across diverse cultures. Baseline evaluations reveal that while there is substantial variation across languages, current models demonstrate a consistent "literal-over-idiomatic" advantage. Also, the results underscore that models struggle with robust idiom understanding without contextual cues. Future work aims to expand the dataset and integrate new factors that influence processing, such as abstractness and imageability.

To protect the integrity of the benchmark against "data contamination" (LLMs training on the test questions), we are implementing a controlled release. A subset of the data was utilized for the AdMIRe 2.0 shared task in 2026 (Arslan et al.,

2026). After the initial one year period, the dataset will be available to the broader research community upon request, provided they agree to terms that prohibit using the data for public LLM training.

References

Doğukan Arslan, Hüseyin Anıl Çakmak, Gülşen Eryiğit, and Joakim Nivre. 2025. Using LLMs to advance idiom corpus construction. In *Proceedings of the 21st Workshop on Multiword Expressions (MWE 2025)*, pages 21–31, Albuquerque, New Mexico, U.S.A. Association for Computational Linguistics.

Doğukan Arslan, Rodrigo Wilkens, Wei He, Dilara Torunoglu Selamet, Thomas Pickard, Aline Villavicencio, Adriana Silvina Pagano, and Gülşen Eryiğit. 2026. MWE-2026 shared task: AdMIRe 2 advancing multimodal idiomaticity representation. In *Proceedings of the 22nd Workshop on Multiword Expressions (MWE 2026)*, pages 276–287, Rabat, Marocco. Association for Computational Linguistics.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction artifacts in metaphor identification datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP 2023)*, pages 6581–6590, Singapore. Association for Computational Linguistics. Main Conference.

Paul Cook, Afsaneh Fazly, and Suzanne Stevenson. 2008. The VNC-tokens dataset. *Towards a Shared Task for Multiword Expressions (MWE 2008)*, page 19.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2025. Investigating idiomaticity in word representations. *Computational Linguistics*, 51(2):505–555.

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.

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. 2025. Semeval-2025 task 1: Admire – advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2597–2609, Vienna, Austria. Association for Computational Linguistics.

Arkadiy Saakyan, Shreyas Kulkarni, Tuhin Chakrabarty, and Smaranda Muresan. 2025. Understanding figurative meaning through explainable visual entailment. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1–23, Albuquerque, New Mexico. Association for Computational Linguistics.

Agata Savary, Manfred Sailer, Yannick Parmentier, Michael Rosner, Victoria Rosén, Adam Przepiórkowski, Cvetana Krstev, Veronika Vincze, Beata Wójtowicz, Gyri Smørdal Losnegaard, Carla Parra Escartín, Jakub Waszczuk, Mathieu Constant, Petya Osenova, and Federico Sangati. 2015. PARSEME – PARSing and Multiword Expressions within a European multilingual network. In *7th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics (LTC 2015)*, Poznań, Poland.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Tongguan Wang, Mingmin Wu, Guixin Su, Dongyu Su, Yuxue Hu, Zhongqiang Huang, and Ying Sha. 2025. MChIRC: A multimodal benchmark for Chinese idiom reading comprehension. In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.